UNIVERSITI TEKNOLOGI MARA

EXPLORING TIKTOK USER-GENERATED CONTENT FOR BUSINESS INTELLIGENCE USING ENHANCED DATA, INFORMATION, KNOWLEDGE AND WISDOM (DIKW) FRAMEWORK

MUHAMMAD AKMAL HAKIM BIN AHMAD ASMAWI

MSc

February 2026

UNIVERSITI TEKNOLOGI MARA

EXPLORING TIKTOK USER-GENERATED CONTENT FOR BUSINESS INTELLIGENCE USING ENHANCED DATA, INFORMATION, KNOWLEDGE AND WISDOM (DIKW) FRAMEWORK

MUHAMMAD AKMAL HAKIM BIN AHMAD ASMAWI

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Information Technology)

Faculty of Computer and Mathematical Sciences

February 2026

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 3 December 2025 to conduct the final examination of Muhammad Akmal Hakim bin Ahmad Asmawi on his Master of Science thesis entitled "Exploring TikTok User-Generated Content For Business Intelligence Using Enhanced Data, Information, Knowledge And Wisdom (DIKW) Framework" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiner recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Mohd Agos Salim Nasir, PhD
Associate Professor
Faculty Of Computer And Mathematical Sciences
Universiti Teknologi MARA
(Chairman)

Kamalia Azma Kamaruddin, PhD
Senior Lecturer
Faculty Of Computer And Mathematical Sciences
Universiti Teknologi MARA
(Internal Examiner)

Nur Haryani Zakaria, PhD
Associate Professor
College of Arts and Sciences
Universiti Utara Malaysia
(External Examiner)

PROFSSOR DR HJH ZURAEDA IBRAHIM
Dean
Institute of Postgraduates Studies
Universiti Teknologi MARA

Date: 26 February 2026

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Muhammad Akmal Hakim Bin Ahmad Asmawi
Student ID. No.	:	2024655976
Programme	:	Master of Science (Information Technology) – CDCS751
Faculty	:	Computer and Mathematical Sciences
Thesis Title	:	Exploring TikTok User-Generated Content For Business Intelligence Using Enhanced Data, Information, Knowledge And Wisdom (DIKW) Framework
Signature of Student	:	 ……..
Date	:	February 2026

ABSTRACT

This research aims to develop a systematic pipeline for transforming TikTok User-Generated Content (UGC) into actionable Business Intelligence (BI) for brands within Malaysia's beauty and personal care sector. While TikTok has become a primary hub for consumer expression, its unstructured and linguistically complex comments characterized by Malaysian slang and code-switching hinder the extraction of reliable insights. To address this, the study adopts the Data-Information-Knowledge-Wisdom (DIKW) hierarchy and the Data Science Trajectories (DST) methodology as its core frameworks. The DST framework guided the operational phases of data acquisition, preparation, and modeling, while the DIKW hierarchy provided the conceptual structure to elevate raw data (comments) into strategic insights (Wisdom). Data were harvested from 20 top-revenue Malaysian influencers, comprising 3,912 videos and 34,597 comments. Preprocessing utilized GPT-4o to normalize informal language, which was then processed through a sentiment analysis model and BERTopic for thematic clustering. Coherence metrics (c_v = 0.620, c_uci = 0.918) confirmed the robustness of the identified topics: Product Usage/Experience and Product Features. Findings reveal a "Revenue Paradox", where high follower counts do not directly correlate with sales; instead, "neutral" comments (24.28%), which often contain specific product inquiries, were identified as having the highest conversion potential. These findings fulfil the research aims by demonstrating that structured analytics can bridge the gap between social media noise and strategic decision-making. The results were consolidated into a Streamlit-based dashboard, which achieved a System Usability Scale (SUS) score of 70.8. This research contributes a validated, modular NLP pipeline that allows businesses to move beyond descriptive metrics toward a deeper, data-driven understanding of consumer behavior in the digital economy.

ACKNOWLEDGEMENT

Alhamdulillah, all praise is due to Allah SWT, the Most Gracious and Most Merciful, for His infinite blessings, strength, and guidance. I am profoundly grateful for the patience and resilience bestowed upon me during the most challenging moments of my research.

I wish to express my deepest and most sincere gratitude to my supervisor, Ts. Dr. Pradeep Isawasan. I am truly humbled by the level of support I have received. Beyond his academic guidance and intellectual insights, they provided the essential financial resources that allowed me to focus entirely on my studies. Most importantly, without their constant encouragement and timely assistance, I would not have been able to complete this research within a year. I am profoundly grateful for his mentorship, which has been a masterclass in both professional excellence and human kindness.

To my beloved parents, Ahmad Asmawi and Haslina, there are no words sufficient to thank you for the lifetime of love and sacrifice that led me to this moment. You have been my steady anchor through every moment of doubt. Your prayers and the quiet strength you provided gave me the resilience to keep going when things felt difficult. Any milestone I achieve is a reflection of your support, and I hope this work brings some happiness to your hearts.

This journey was demanding, but I am acutely aware that I did not walk it alone. I am truly blessed to have had such incredible people standing behind me. Alhamdulillah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Research Background

In today's business landscape, understanding consumer preferences is essential to maintain a competitive edge, particularly in a rapidly evolving marketplace characterized by fluctuating customer demands and emerging industry trends. Companies increasingly rely on various data streams to facilitate informed decision-making and developed adaptive strategies. One of the most valuable data sources is user-generated content (UGC), which offers direct insights into consumer sentiments regarding products and services. User-generated content (UGC) is defined as any form of content, such as text, videos, images, and reviews, that is created and shared by individuals rather than by brands or official representatives (Santos, 2022). This type of content is often posted on social media platforms and other online spaces where users express their thoughts, opinions, and experiences. Example of UGC includes comments, reviews, posts, and videos shared across digital platforms which could provide rich and unfiltered perspectives on customer attitudes, preferences, and behaviours (Djafarova & Rushworth, 2017; Marti-Ochoa et al., 2024).

The academic utility of UGC has been widely demonstrated across various sectors as a superior alternative to traditional market research (Naab & Sehl, 2017). For instance, researchers in the tourism industry have analyzed UGC from travel forums to identify service gaps that impact guest loyalty, while those in the retail sector have used visual UGC to track real-time aesthetic trends. Furthermore, studies on electronic word-of-mouth (eWOM) highlight that consumers often perceive UGC as more credible and trustworthy than brand-sponsored content, making it a critical asset for capturing authentic brand perceptions (Ye et al., 2011).

The rise of social media has fundamentally transformed consumer interaction and communication, implanting itself deeply into people's everyday lives (Bossetta, 2018; Stieglitz et al., 2018). As of January 2024, there are approximately 5.04 billion social media users worldwide, with Malaysia contributing a total of 28.68 million users which represents 83% of the country's population (Kemp, 2024). Figure 1.1 below illustrates the total users for each social media platforms in Malaysia. Based on the

figure, it shows that all social media users in Malaysia have registered for TikTok accounts, in an equal number of social media users. Following TikTok are YouTube and, with 24.1 and 22.35 millions users respectively. These platforms have evolved as a vital source of consumer knowledge rather than just being channels for advertising (Bossetta, 2018). Platforms such as TikTok, YouTube, and Facebook enable businesses to access extensive audiences and serve direct engagement. Businesses can leverage the significant volume of UGC resulting from such engagement to understand brand perception and customer preferences, thereby fine-tuning their marketing strategies accordingly.

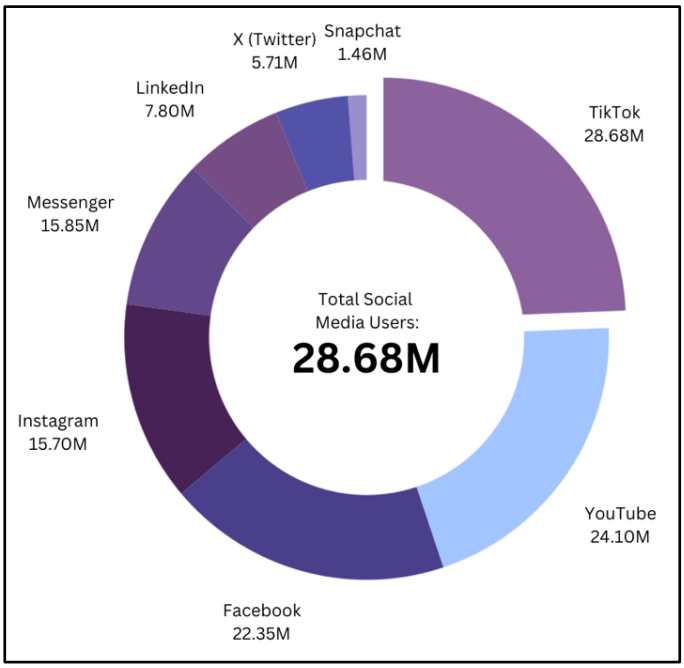


Figure 1.1 The Total Of Social Media Users In Malaysia

(Source: (Kemp, 2024))

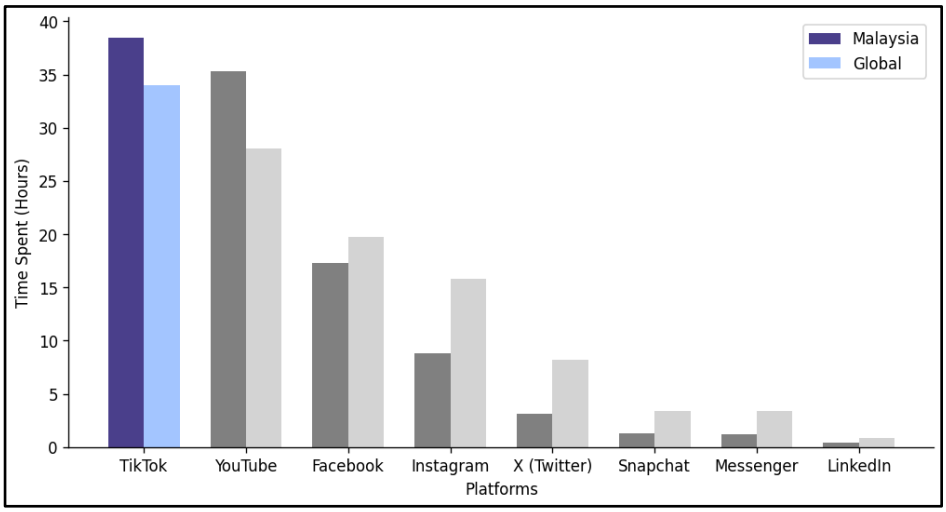


Figure 1.2 The Amount Of Time Spent On Each Platforms

(Source: (Kemp, 2024))

Figure 1.2 illustrates the amount of time spent on social media platforms in which shows how much social media integrate into the daily routines of its users. On a global scale, user spends approximately 34 hours per month on TikTok, 28 hours on YouTube, and 19 hours on Facebook. In Malaysia, these figures are even more noticeable, with monthly usage nearly 40 hours on TikTok, 35 hours on YouTube, and 17 hours on Facebook. These high levels of engagement highlight the huge potential for businesses to extract consumer insights from these platforms, particularly given the vast volume of UGC being produced on an ongoing basis.

Furthermore, there has been a notable increase in eCommerce activities, with 56.1% of global internet users are engaged in online shopping activities every week. In Malaysia, this figure is even higher at 61.9%, underscoring the increasingly critical role of digital platforms in influencing consumer's purchasing behaviour. Social media has become a major factor in the discovery of new brands and products. In Malaysia, 40.1% of internet users report that they discovers new brands, products, and services through social media advertisements, positioning platforms like TikTok, YouTube, and Facebook as essential tools to enhance brand visibility and building consumer engagement (Gutierrez et al., 2023).

TikTok, specifically, has experienced remarkable popularity due to its distinctive features that encourage creative expression and interactive user engagement. The TikTok algorithm plays a critical role in content visibility, leveraging factors such as user interactions, video information (including captions, sounds, and hashtags) to

personalize the content shown to each user, making TikTok a vibrant source of UGC (Bhandari & Bimo, 2020). However, this unstructured and dynamic nature also poses significant challenges for data analysis (Stieglitz et al., 2018). The variety forms of medium (such as videos, comments, and captions) with diverse styles and expressions, adds the complexity to make a systematic data collection, robust data pre-processing, and insight's interpretation. Nonetheless, effectively harnessing this data can yield highly authentic consumer insights, thus offering a more nuanced understanding of brand-customer interactions.

This research primarily aimed to generate actionable business insights from TikTok user-generated content in the Beauty and Personal Care category. The unstructured and evolving nature of TikTok data required moving beyond straightforward, linear methodologies, which are often unsuitable for modern, repetitive, and dynamic problems. To address this, the study enhanced existing approaches with a more flexible and iterative framework capable of handling bilingual comments, slang, and multimedia formats. Within this framework, advanced analytical techniques such as BERTopic for topic modeling and GPT-based sentiment analysis were applied to uncover key themes and consumer sentiments. The findings were then presented through a user-friendly dashboard, enabling businesses to interpret trends, topics, and perceptions in a clear and practical way that supports strategic decision-making.

1.2 Problem Statement

The rise of user-generated content (UGC) on TikTok presents businesses with valuable opportunities to understand audience behavior and track emerging trends. Unlike text-heavy platforms like X (Twitter) or Facebook, TikTok relies on short videos and interactive comments, creating a unique but complex data landscape. However, extracting reliable and actionable insights from this diverse content remains a major challenge (Stieglitz et al., 2018). The core issue is that businesses struggle to make sense of TikTok's unstructured and evolving content, which limits their ability to track consumer sentiment, identify engagement drivers, and support data-driven decision-making. Figure 1.3 illustrates the structured flow of the research problem, showing how the underutilization of TikTok UGC breaks down into specific challenges, their root causes, and the resulting impact on businesses. The solutions developed in this research

address these challenges, enabling the extraction of business insights that can inform strategy and enhance competitiveness.

The first major challenge in analyzing TikTok data lies in the complexity of its content structure and the inconsistency of data quality. TikTok's combination of videos, text-based interactions, and varied metadata creates a highly unstructured landscape. Unlike platforms such as X (Twitter) or Facebook, where textual analysis is more straightforward, TikTok's mixed content format makes integration and analysis far more difficult. This complexity means that businesses often struggle to derive consistent insights without tailored approaches that can handle diverse and evolving data types. A closely related challenge is the quality of TikTok comments, which frequently include slang, abbreviations, emojis, and local dialects. These informal patterns are difficult for standard NLP techniques to process, often resulting in inaccurate sentiment detection and misinterpretations. When such content is not carefully prepared, businesses risk relying on misleading insights that weaken strategic decisions. Addressing these issues requires effective data preprocessing and structured handling of TikTok content so that advanced analytics can produce results that are both reliable and actionable.

The lack of structured and adaptable methodologies for analyzing TikTok data has limits its full potential as a business intelligence resource, primarily due to the challenges in processing unstructured, multimodal, and slang-heavy content (Stieglitz et al., 2018). Existing approaches often fail to provide a systematic workflow for handling TikTok's evolving content and linguistic diversity, which affects key stages such as data collection, preprocessing, modeling, and evaluation. As a result, advanced techniques such as BERTopic for topic modeling and GPT-based sentiment analysis struggle to deliver accurate results, since they depend on properly prepared and contextually reliable data. Without flexible methods to guide their application, these tools are likely to produce inconsistent findings, leading to unreliable insights and misinterpretations that weaken business decisions.

In conclusion, addressing TikTok's content complexity, data quality issues, and methodological limitations is critical for unlocking its value as a source of business intelligence. This research tackled these challenges by enhancing existing methodologies with a more iterative and adaptable framework, integrating robust data preprocessing with advanced NLP and analytics. These improvements made it possible to extract reliable and actionable insights from TikTok's raw data, enabling businesses

to transform unstructured content into strategic intelligence that supports decision-making and strengthens competitiveness.

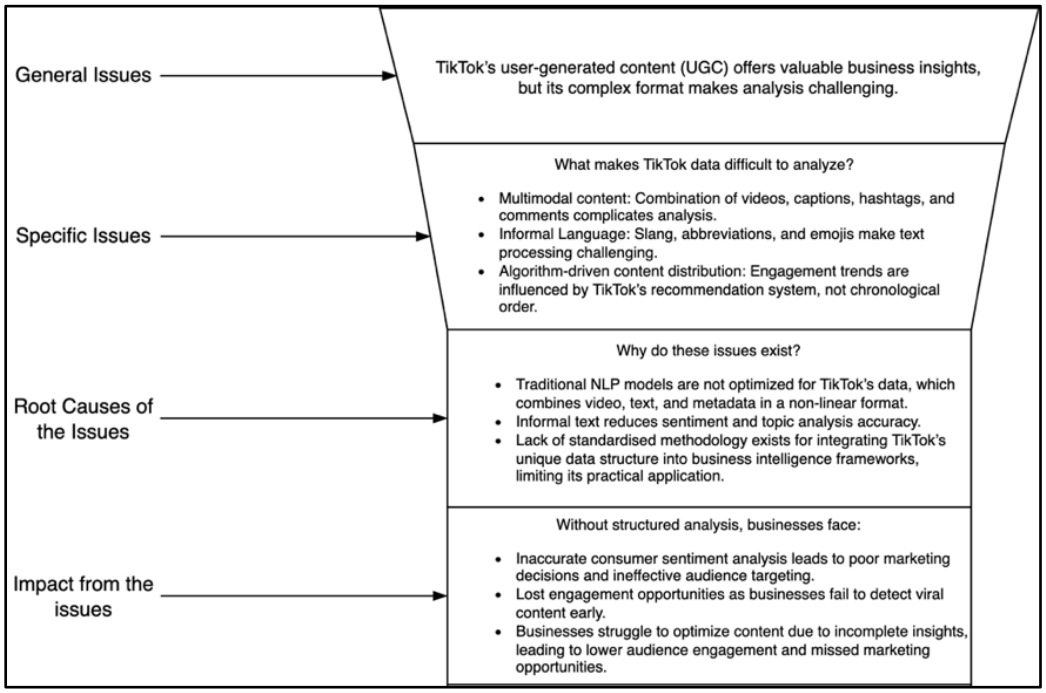


Figure 1.3 The Research Problem Flow

1.3 Research Question

The following are the research questions for this research:

RQ1) What are the essential components of an adaptable methodology for systematically processing TikTok's unstructured data to support reliable insights?

RQ2) Which advanced analytics techniques can be applied to TikTok descriptive metadata to uncover patterns, trends, and engagement factors?

RQ3) How can topic modeling and sentiment analysis techniques, including BERTopic and GPT, be applied and evaluated to derive actionable insights from TikTok comments?

RQ4) How can actionable business insights from TikTok data be extracted and effectively visualized through a user-friendly dashboard?

1.4 Research Objectives

The following are the objectives for this research:

RO1) To establish a structured and adaptable methodology for processing TikTok's unstructured data for reliable insight extraction.

RO2) To analyze TikTok descriptive metadata to identify patterns, trends, and engagement factors that influence audience behavior.

RO3) To apply advanced modeling techniques, including BERTopic for topic modeling and GPT-based sentiment analysis, to generate meaningful insights from TikTok comments.

RO4) To extract actionable business insights from TikTok data and visualise through a user-friendly dashboard for data-driven decision-making.

1.5 Research Scope

This research applied Information Technology (IT) methodologies to structure, analyze, and visualize social media data for business intelligence decision-making, focusing on the application of existing algorithms rather than their development. The study examined TikTok user-generated content in the Beauty and Personal Care category, which was selected due to its status as the highest revenue-generating sector on the platform and also due to its critical relevance to brands and businesses in Malaysia, which increasingly rely on social commerce for market penetration and growth, aiming to transform unstructured data into actionable business insights. Consequently, the scope of the research covered postings from the top 20 Malaysian influencers in this category, who were purposively chosen based on their highest generated revenue to ensure the analysis focused on commercially impactful content. The dataset included comments written in both Malay and English, collected between January and August 2024. The research includes the analysis of textual content (comments) and descriptive metadata (likes, shares, views). It excludes the processing of video frames (pixel analysis) or audio data, as the focus is on sentiment and topic modeling rather than computer vision or audio signal processing. A review of existing social media analytics literature was first conducted to identify gaps and limitations in

current approaches. Based on these findings, a structured yet adaptable framework was developed to process TikTok data more effectively.

A pipeline was established to collect and clean TikTok metadata and comments, addressing challenges such as multilingual text, slang, abbreviations, and the informal nature of short comments. Advanced techniques, including BERTopic for topic modeling and GPT-based sentiment analysis, were then applied to uncover key themes and consumer sentiments. Finally, the findings were presented through an interactive dashboard, which visualizes trends, topics, and sentiments in a clear and accessible format. This dashboard served to make the results interpretable and actionable, ensuring that the insights could directly support data-driven decision-making.

1.6 Significance of Research

This research addressed key gaps in the use of social media for business intelligence by focusing specifically on TikTok, a fast-growing but under-researched platform. Social media continues to evolve rapidly, shaping how consumers behave and perceive brands. This research closed the gap between existing theoretical approaches and practical tools by transforming user-generated content (UGC) into business insights that are both reliable and actionable. The main contributions of this research are in several areas. First, it generated actionable business insights by applying advanced analytical techniques such as BERTopic for topic modeling and GPT-based sentiment analysis. These methods revealed key discussion themes, emerging trends, and sentiment shifts in consumer conversations, showing what resonates with audiences and how perceptions change over time. Such insights help businesses understand consumer interests, brand reputation, and common concerns, enabling data-driven adjustments to marketing and engagement strategies. A central output was the development of an interactive dashboard that visualizes trends, topics, and sentiments in real time. This makes complex findings easy to interpret and apply, supporting managers in making timely and informed decisions.

Second, the research contributed to data quality and processing by establishing a robust pipeline for collecting and cleaning TikTok metadata and comments. Addressing issues such as slang, abbreviations, emojis, and bilingual text directly improved the reliability of the insights, overcoming a common weakness in earlier studies. Third, the research added to academic knowledge and methodology by

reviewing existing literature, identifying flaws in current approaches, and enhancing them with a more structured and adaptable framework suited to modern, fast-changing data environments. Finally, by focusing specifically on TikTok, this research sets itself apart from prior work that has concentrated on platforms such as Twitter and Facebook. TikTok's rapid growth and strong influence on consumer behavior make it an important yet underutilized data source. Overall, the research advances both academic understanding and practical business applications by showing how TikTok UGC can be transformed into strategic intelligence. The ability to track shifting consumer sentiment, identify emerging topics, and monitor brand perception in real time gives businesses a powerful tool to adapt quickly to market dynamics and optimize their content strategies.

1.7 Relevance to Information Technology

This research contributes to the field of Information Technology (IT) by demonstrating how unstructured social media data can be transformed into structured insights that support business intelligence. While it employed natural language processing (NLP) and machine learning techniques, the primary contribution lies in the areas of data engineering, big data analytics, and decision support systems (DSS). Through RO1 and RO2, the study applied a structured methodology and data pipeline to process TikTok metadata and comments, contributing to IT-based approaches for data management and information retrieval. RO3 and RO4 extended this by focusing on business intelligence and decision support, applying BERTopic for topic modeling and GPT-based sentiment analysis. The resulting insights were delivered through an interactive dashboard, allowing businesses to leverage IT tools for market analysis and digital strategy. By integrating IT principles into data-driven content strategy and social media analytics, this research advanced IT's role in information retrieval, trend detection, and enterprise decision support, reinforcing its relevance within applied IT and business intelligence solutions.

1.8 Thesis Outline

This research consist of eight chapters. Chapter 1 presents a general outlook on TikTok as a new user-generated content (UGC) platform and discusses how business intelligence might benefit from it. This chapter begins with the research background,

problem statement, research objectives, research questions, the research scope, research significance, and thesis outline.

Chapter 2 reviews existing research on social media and business intelligence through a bibliometric and text analytics approach. It explains how literature data was collected and analyzed, highlighting trends, influential studies, and emerging themes in the field. The chapter discusses the evolution of social media, business intelligence, and user-generated content, examining their intersections and research gaps. It further explores platform comparisons, analytical techniques, and specialized applications across industries. Finally, this chapter introduces the DIKW framework and the Data Science Trajectories (DST) model as guiding methodologies, setting the foundation for the thesis's research design.

Chapter 3 introduces the research's methodological framework, structured around the Data Science Trajectories (DST) model to analyze TikTok influencer activity in the beauty and personal care sector. It outlines the business context, identifies TikTok as the primary data source, and explains the acquisition of revenue data from Kalodata and video or comment data from Apify. The chapter details the variables considered, the data preparation process and highlights the integration of datasets to support descriptive analytics, sentiment analysis, and topic modeling. It also provides a cost breakdown and concludes by positioning of the methodology as the foundation for the analyses presented in later chapters.

Chapter 4 applies descriptive analytics to the prepared datasets, focusing on revenue data and TikTok video metadata. It begins with exploratory data analysis for general patterns in revenue, followers, views, and engagement, then moving into detailed revenue analytics, including summary statistics, correlation analysis, and top-performer comparisons. Metadata analytics assess video-level engagement metrics, hashtag usage, and temporal patterns, highlighting how content quality, length, and timing influence audience interaction. The chapter concludes by showing that diversified revenue streams, optimized hashtag strategies, and culturally timed content drive stronger engagement and monetization outcomes.

Chapter 5 applies GPT-4o to classify 34,597 TikTok comments from Beauty and Personal Care influencers into Positive, Neutral, and Negative categories. After cleaning and standardizing the data, the model revealed the sentiment distributions. Analyses were conducted at overall, word-level, and influencer-specific scales, highlighting the strong sentiment indicators and showing the revenue correlations with

sentiment labels. The chapter concludes showing several key conversion opportunity, effective monetization, and diversified strategies.

Chapter 6 focuses on the application of topic modeling to analyze TikTok comments from the Beauty and Personal Care category. The chapter outlines the full modeling pipeline, starting with text cleaning and embedding creation, followed by dimensionality reduction, clustering, and topic generation with BERTopic. Coherence measures such as c_v, u_mass, c_uci, and c_npmi are used to evaluate topic quality. Visualizations are included to illustrate topic distributions, keywords, and example comments. The chapter concludes with a discussion on the interpretability and business relevance of the identified topics.

Chapter 7 presents the dashboard implementation, which serves as the output of the research. It explains the system architecture, integrating the data pipeline, GPT-based sentiment analysis, and BERTopic modeling into a Streamlit dashboard. The chapter outlines dashboard design objectives, focusing on clarity, usability, and decision support, and then details the main dashboard components, including influencer metrics, sentiment trends, topic modeling insights, and business intelligence visualizations. Interactivity and user experience are discussed in terms of navigation, filtering, and responsiveness, showing how the dashboard supports both analysis and managerial decision-making. Usability testing results are also presented, confirming the dashboard's effectiveness while identifying areas for refinement. The chapter concludes by highlighting the dashboard's role as a bridge between technical outputs and actionable business insights.

Chapter 8 synthesizes the findings of the research by revisiting the research objectives and highlighting how each was addressed across previous chapters. It consolidates the conceptual, technical, and empirical contributions, showing how the constructed pipeline supports decision-making in the Beauty and Personal Care domain on TikTok. The chapter also discusses the broader implications for Information Technology, particularly in translating unstructured social media data into structured, actionable insights. Finally, it reflects on the research's limitations and proposes directions for future research, setting the stage for continued refinement and application of the methodology in wider contexts.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter provides a detailed overview of research on social media and business intelligence through a bibliometric analysis. It starts by explaining how the data was collected and the methods used for analysis. The focus is on finding key trends, important authors, and influential research in this area. The chapter uses different tools, like exploratory data analysis (EDA), citation analysis, and author productivity analysis, to track how research has grown and its impact over time. It also looks at text analytics, applying techniques such as word clouds, keyword co-occurrence, and topic modeling. These methods help reveal the main themes and research directions in the field. By doing so, the chapter offers insights into how research on social media and business intelligence has evolved and where it might go in the future.

2.2 Methodology of Literature Review

This section explains how the literature for this research was collected and analyzed. The focus is on using bibliometric analysis to understand research trends, key themes and research gaps in the field of social media and business intelligence. The steps taken to perform this analysis and the tools used are described below.

2.2.1 Bibliometric Analysis

Bibliometric analysis is a systematic method for evaluating the structure, output, and impact of academic research (Donthu et al., 2021). It quantitatively examines publication trends, citation patterns, and keyword co-occurrences, providing insights into the evolution of research fields. This method is particularly valuable in interdisciplinary domains like social media and business intelligence (BI), where research is rapidly expanding. By focusing on measurable outputs, bibliometric analysis reduces manual bias and ensures an objective assessment of the academic landscape (Donthu et al., 2021; W. Zhang, Yang, et al., 2023). In this research, bibliometric analysis systematically reviews a large body of literature, identifying key research

trends, influential studies, and evolving topics in social media and BI. By analyzing citation patterns and keyword relationships, this approach uncovers emerging research areas, including multilingual social media analytics and real-time data processing, which remain underexplored. These insights help frame thematic clusters and highlight research gaps.

Compared to traditional literature reviews, bibliometric analysis minimizes subjective bias by relying on quantitative metrics such as citation counts, keyword clustering, and research impact indicators (Öztürk et al., 2024). Furthermore, while both Bibliometric Analysis and Systematic Literature Review (SLR) are formal and reproducible methods, they serve distinct purposes. An SLR typically addresses a narrow and specific research question by synthesizing a selective group of high-quality studies. In contrast, bibliometric analysis is chosen for this research to provide a broad and quantitative mapping of the entire scientific field. While an SLR provides depth into what researchers found, bibliometric analysis provides the breadth needed to identify how the field has evolved and to map thematic clusters across a large body of literature that a more restrictive SLR might overlook. It also offers visual representations of author productivity, thematic clusters, and evolving research trajectories, enabling researchers to quickly identify high-impact studies and knowledge gaps (Donthu et al., 2021). Common bibliometric techniques include co-citation analysis, bibliographic coupling, and keyword clustering, which reveal patterns in research focus and collaboration networks. Despite its strengths, bibliometric analysis has limitations. Citation-based metrics often prioritize well-established studies, potentially overlooking newer but impactful contributions. Additionally, database coverage can influence results, as certain fields and emerging research areas may be underrepresented (Donthu et al., 2021; Lim & Kumar, 2024; Öztürk et al., 2024). Recognizing these challenges ensures a balanced interpretation of bibliometric findings.

In this research, bibliometric analysis forms the foundation for understanding research trends and gaps in social media and BI. It complements text analytics by structuring key themes, refining topic modeling, and enhancing thematic categorization. Together, these methods provide a comprehensive, data-driven approach to analyzing the academic landscape and advancing knowledge in this domain.

2.2.2 Data Collection

The data for this research was collected from the Web of Science Core Collection, a database recognized for its extensive indexing of high-quality peer-reviewed academic publications (V. K. Singh et al., 2021; Vlase & Lähdesmäki, 2023). The utilisation of only the Web of Science (WoS) Core Collection as the sole data source was driven by its status as the "gold standard" for high-impact scientific research. Unlike broader databases such as Scopus or Google Scholar, WoS employs rigorous indexing criteria that prioritize high-quality, peer-reviewed journals with established impact factors (V. K. Singh et al., 2021). This selectivity acts as a necessary quality filter for this research, ensuring that the bibliometric analysis is grounded in validated, influential scholarship rather than diluted by predatory or low-quality publications often found in less regulated databases (Pranckutė, 2021; Vlase & Lähdesmäki, 2023). Furthermore, WoS offers superior historical depth (dating back to 1900) compared to Scopus (launched in 2004), making it uniquely suited for tracing the long-term evolution of Business Intelligence from its early theoretical foundations to modern AI applications. While restricting data to a single database may introduce coverage bias, specifically an overrepresentation of English-language and Western-centric journals. This limitation was mitigated by the specific scope of this research. The domains of Business Intelligence (BI) and Information Technology (IT) are predominantly driven by global, English-language discourse in major international journals. Therefore, the "bias" towards high-impact international journals serves the research objective of identifying global best practices and dominant technological trends rather than fragmented, region-specific discussions. To further counter potential coverage gaps, the search strategy employed broad, inclusive keywords (e.g., "social networking site" alongside "social media") to capture a wide array of relevant studies within this high-quality dataset. The collection process was designed to ensure comprehensive coverage of research on social media and business intelligence (BI). The following steps outline the data collection process. The search included broad terms such as "social media", "social networking site", "business intelligence", and "business analytic". It is intentionally to exclude specific platform name such as Twitter or TikTok, as the research wants to capture a comparative discussion. This approach allows for an unbiased identification of the platforms most commonly used for BI research and those that are underexplored. The search was conducted on titles, abstracts,

and keywords to ensure a focused yet inclusive dataset. This targeted strategy minimizes the inclusion of irrelevant studies while capturing the core content of relevant research.

The search was designed to cover a broad time range from 1970 to 2024, focusing on publications that align with this timeframe. This does not imply that publications within the dataset start as early as 1970, however, this range sets the boundaries for capturing research trends over an extended period. The actual dataset reflects the years when relevant research began appearing in the field. After applying the search criteria, a total of 387 documents were retrieved, including journal articles, conference proceedings, and reviews. These documents were considered sufficient to provide a comprehensive overview of the field. This search strategy ensures a comprehensive and unbiased dataset for bibliometric and text analytics, capturing trends, tools, and platforms in social media-driven BI research. Table 2.1 shows the summary of the search criteria.

Table 2.1
Paper Searching Criteria

Parameter	Details
Collection	Web of Science Core Collection
Search Field	Topic (titles, abstracts, keywords)
Search String	(“social media*” OR “social networking site*”) AND (“business intelligence” OR “business analytic*”)
Date Range	All years (1970-2024)
Date	26 August 2024

2.2.3 Data Analysis

The data analysis in this research combines bibliometric methods and advanced text analytics to explore research trends, identify influential contributors, and uncover thematic clusters within the academic literature on social media and business intelligence (BI). This approach consists of two key components which are Literature Analytics and Text Analytics, each addressing distinct aspects of the research landscape.

Literature analytics provides a quantitative overview of social media and BI research by examining publication trends, citation counts, and research domains. This

begins with exploratory data analysis (EDA) to assess publication years, document types, and citation networks. Citation analysis identifies high-impact studies and author influence, while publication trends track research growth over time. Research area analysis maps how social media and BI intersect across academic disciplines, offering insights into evolving research directions. Together, these analyses structure the research landscape and form the foundation for further qualitative exploration. Text analytics delves into semantic content by analyzing titles, abstracts, and keywords. Compared to literature analytics, which focuses on quantitative trends, text analytics extracts deeper thematic insights. The process begins with a word cloud to visualize frequently used terms, followed by a clustered tree map, which groups keywords into thematic categories. A keyword co-occurrence analysis further examines the term "business intelligence", mapping its contextual relationships within the dataset.

After keyword-based exploration, topic modeling extends this analysis by clustering research abstracts into coherent themes (Grootendorst, 2022). The first step involves embedding textual data to capture semantic relationships. This research employs SciBERT, a domain-specific adaptation of BERT, selected for its ability to preserve contextual accuracy in scientific texts (Beltagy et al., 2019). To manage high-dimensional data, Uniform Manifold Approximation and Projection (UMAP) is used for dimensionality reduction, enabling better visual separation of topic distributions (Ghojogh et al., 2021). UMAP was chosen over PCA and t-SNE due to its superior preservation of semantic structures in textual embeddings. Next, HDBSCAN, a density-based clustering algorithm, identifies cohesive research themes. Unlike k-means, HDBSCAN does not require predefining the number of clusters, making it ideal for uncovering naturally occurring topic distributions (Scoccola & Rolle, 2023a). Finally, the research applies BERTopic for topic modeling, which improves interpretability and coherence compared to traditional models like LDA (Devlin et al., 2018; Grootendorst, 2022). The generated topics are further refined using a customized term-weighting method, ensuring the most relevant terms are emphasized for better thematic clarity.

This combination of bibliometric and text analytics techniques provides a comprehensive understanding of the academic landscape. However, challenges remain. While SciBERT embeddings enhance domain-specific accuracy, they may struggle with ambiguous terminology. Additionally, HDBSCAN's sensitivity to hyperparameters can affect topic stability. Despite these limitations, this integrated

approach provides a structured, data-driven framework that reveals emerging research priorities, underexplored themes, and evolving trends in social media and BI analytics.

2.3 Background

2.3.1 Social Media

Social media has undergone a transformative evolution, expanding from early platforms like Friendster and Myspace to complex digital ecosystems such as Facebook, Instagram, X (Twitter), and TikTok. These platforms no longer serve as mere communication tools; instead, they have evolved into multi-functional spaces that integrate real-time interactions, multimedia sharing, and AI-driven content curation. The ability of social media to blur the lines between content consumers and creators has redefined digital participation, enabling users to not only consume content but also actively shape narratives and discussions (González-Bailón & Lelkes, 2023; Zsila & Reyes, 2023). This shift from static, one-way communication to dynamic, participatory engagement has positioned social media as a powerful force in shaping societal discourse, business strategies, and global politics. However, its impact extends far beyond connectivity, influencing both innovation and controversy as it reshapes how people interact, access information, and perceive reality.

Social media networks not only connect users but also shape their behaviours and interactions through algorithmic personalization and predictive analytics. These platforms rely on data-driven recommendation systems to enhance user engagement, ensuring that content is customized based on past interactions, preferences, and browsing patterns. While this personalization improves content relevance, it also reinforces behavioural patterns, potentially leading to filter bubbles where users are exposed primarily to content that aligns with their pre-existing beliefs. One effective approach for analyzing these patterns is Social Network Analysis (SNA), which helps researchers and businesses examine user connections, content dissemination, and influencer impact. Studies indicate that digital engagement follows power-law distributions, where a small fraction of highly active users which typically influencers, content creators, and verified accounts, generate the majority of interactions (I. Ali et al., 2023). This dynamic is especially visible on platforms like X (Twitter) and TikTok,

where high-profile users have disproportionate influence, often shaping trending discussions and public narratives.

From a technological standpoint, social media functions as a techno-social ecosystem, where AI-driven algorithms, human behaviours, and emotional engagement collectively determine information visibility and user experience. These platforms continuously refine their recommendation systems, prioritizing content that maximizes user engagement and retention. While such mechanisms optimize content delivery, they also contribute to the amplification of emotionally charged material, often favouring sensationalism, outrage, and polarizing viewpoints. Research highlights how affective content reinforcement plays a crucial role in public discourse and social polarization, as emotionally driven posts tend to generate more interactions, regardless of factual accuracy (Steinert et al., 2025). The result is an ecosystem where high-arousal content dominates, fostering an environment where critical reflection is often overshadowed by immediate emotional reactions.

The behavioural impact of social media is multifaceted, influencing both positive and negative user experiences. On one hand, these platforms facilitate advocacy, public awareness, and crisis response, enabling global movements, public health campaigns, and real-time disaster relief coordination. For instance, during the COVID-19 pandemic, social media served as a critical information hub, allowing health organizations to disseminate updates, counter misinformation, and engage with the public. However, social media also introduces behavioural risks, particularly in the form of compulsive usage, algorithmic content reinforcement, and the psychological effects of digital validation. Young users, particularly on visually immersive platforms like TikTok and Instagram, are especially vulnerable to the impact of curated, idealized imagery, which can influence body dissatisfaction, self-objectification, and mental health concerns (Harriger et al., 2023). These concerns have led to the development of media literacy initiatives such as Social Media Literacy (SoMeLit), which aim to promote critical digital consumption, enhance user awareness, and mitigate potential harms associated with algorithmic-driven content exposure (H. Cho et al., 2024).

Beyond its cultural and psychological influence, social media has also transformed marketing, business intelligence, and consumer engagement. Platforms offer hyper-personalized advertising, predictive analytics, and real-time sentiment tracking, allowing brands to target specific demographics with unparalleled precision. Businesses leverage social media listening tools to analyze customer sentiment, market

trends, and competitor strategies, enabling data-driven decision-making and campaign optimization. However, these advancements also raise ethical concerns, particularly regarding data privacy, user profiling, and algorithmic biases in content exposure. Companies must navigate the fine line between leveraging consumer insights for marketing strategies and ensuring ethical data usage practices. The evolution of social media has fundamentally reshaped communication, commerce, and consumer behavior, offering unprecedented access to real-time insights while simultaneously raising concerns about misinformation, ethical AI applications, and digital well-being. As these platforms continue to evolve, the challenge will be balancing technological innovation with responsible engagement practices. Researchers, policymakers, and businesses must work toward developing ethical guidelines, improving transparency in algorithmic processes, and fostering a digital landscape that prioritizes both user experience and societal well-being.

As social media platforms continue to evolve, they generate vast amounts of user-generated data, engagement metrics, and sentiment trends, which businesses can leverage for data-driven decision-making. However, the sheer volume and unstructured nature of this data present significant challenges in extracting meaningful insights. To address these challenges, organizations increasingly rely on Business Intelligence (BI), a field that integrates data management, analytics, and visualization tools to transform raw social media data into actionable insights. By applying BI techniques to social media analytics, businesses can monitor consumer behavior, market trends, and brand perception in real time, allowing for more informed strategic decisions.

2.3.2 Business Intelligence

Business Intelligence (BI) has evolved from its early use in Decision Support Systems (DSS) and Relational Database Management Systems (RDBMS) into a modern approach that helps businesses turn raw data into meaningful insights. BI development began in the 1970s with RDBMS, which introduced structured data management using SQL-based queries and ACID (Atomicity, Consistency, Isolation, Durability) compliance to ensure data integrity and reliability. In the 1990s, BI expanded with data warehousing and Online Analytical Processing (OLAP) tools, which allowed businesses to store and analyze large volumes of structured data. By the 2000s, new database technologies, such as NoSQL, helped companies handle unstructured data at

scale, and in-memory databases like SAP HANA provided real-time analytics by storing data directly in memory instead of on traditional disk storage. In recent years, BI has integrated machine learning (ML) and artificial intelligence (AI), allowing businesses of all sizes, including small and medium-sized enterprises (SMEs), to make faster and more accurate data-driven decisions (Praful Bharadiya, 2023; Shahadat Hosen et al., 2024). BI systems consist of several key components, including data storage, data processing, analytics tools, and reporting frameworks. Data management plays a fundamental role in ensuring that information is accurate, consistent, and accessible. Traditional RDBMS provides structured data reliability, while NoSQL databases offer scalability for handling semi-structured and unstructured data. NewSQL databases combine features of both systems, making them ideal for handling large transactional workloads, such as those in e-commerce platforms (Praful Bharadiya, 2023; Shahadat Hosen et al., 2024). To process and clean data, businesses use Extract, Transform, Load (ETL) frameworks, which help ensure that data is consistent, structured, and ready for analysis (Bharadiya, 2023; Shahadat Hosen et al., 2024).

BI systems use different types of analytics to process and interpret data. Descriptive analytics looks at past data to explain historical trends and patterns. Predictive analytics uses statistical models and machine learning algorithms to forecast future trends, helping businesses anticipate customer behavior, market trends, and risks. Prescriptive analytics goes further by providing actionable recommendations, allowing businesses to optimize decision-making through scenario simulations and strategic planning (Bharadiya, 2023; Tsiu et al., 2024). Data visualization tools such as Tableau and Power BI help businesses interpret complex data by creating interactive dashboards and reports that allow users to make informed decisions quickly (Praful Bharadiya, 2023; Shahadat Hosen et al., 2024; S. Singh et al., 2023). In addition to analytics tools, businesses use frameworks like the Technology-Organization-Environment (TOE) model to ensure that they are ready to adopt BI technologies, while Explainable AI (XAI) frameworks help improve trust and transparency in AI-driven BI systems (Shahadat Hosen et al., 2024; Tsiu et al., 2024).

BI is widely used across many industries to improve efficiency and decision-making. In retail, BI helps businesses set pricing strategies, manage inventory, and create targeted marketing campaigns. Businesses use clustering techniques to group customers by behavior, making personalized marketing more effective (Bharadiya, 2023; Hamzehi & Hosseini, 2022; Tsiu et al., 2024). In healthcare, BI improves patient

care, resource management, and early disease prediction, allowing healthcare providers to identify high-risk patients and optimize treatments (Emmanuel Osamuyimen Eboigbe et al., 2023; Rehman et al., 2022; Shahadat Hosen et al., 2024). In financial services, BI enhances fraud detection and risk management by using machine learning to identify unusual patterns in transactions, helping banks detect fraudulent activities in real time (Bharadiya, 2023; Shahadat Hosen et al., 2024). In higher education, BI supports student performance tracking, retention analysis, and institutional planning, helping universities make data-driven decisions about admissions and course planning (Hmoud et al., 2023). In manufacturing and supply chain management, BI tools improve demand forecasting, supplier performance monitoring, and operational efficiency, ensuring smoother production and logistics management (Al-Okaily et al., 2023; Jahani et al., 2023; Shahadat Hosen et al., 2024).

Despite its many benefits, BI implementation faces several challenges. One major issue is data quality. If data is incomplete, inconsistent, or inaccurate, it can lead to incorrect conclusions and poor decision-making. Businesses must invest in data cleaning and governance frameworks to maintain high-quality, reliable data (Al-Okaily et al., 2023; Bharadiya, 2023; Shahadat Hosen et al., 2024). Another challenge is the complexity of BI tools, and the skills required to use them effectively. Many organizations, especially SMEs, struggle to adopt BI due to a lack of trained personnel and the technical expertise needed to operate advanced BI systems (Shahadat Hosen et al., 2024; Tsiu et al., 2024). Organizational resistance is another barrier. Employees may be reluctant to shift from intuition-based decision-making to data-driven processes, leading to slow adoption and underutilization of BI tools (Shahadat Hosen et al., 2024; Tsiu et al., 2024). Ethical and security concerns are also significant. AI-driven BI systems can introduce algorithmic bias, leading to unfair decision-making in areas like hiring, loan approvals, and insurance claims. Moreover, businesses must comply with data privacy laws like GDPR and CCPA to protect sensitive user data and prevent unauthorized access or misuse (Bharadiya, 2023; Shahadat Hosen et al., 2024).

To fully unlock the potential of BI, organizations must address both technical and organizational barriers. Investing in data governance, employee training, and ethical AI frameworks can help businesses make better use of BI for innovation, strategic planning, and operational efficiency. As BI continues to evolve, integrating AI-powered analytics, real-time data processing, and transparent decision-making

frameworks will be essential for organizations to remain competitive in an increasingly data-driven world.

2.3.3 User Generated Content (UGC)

User-generated content (UGC) refers to digital media created by individuals rather than organizations, encompassing text, images, videos, and interactive contributions. It has become a fundamental element of digital ecosystems, shaping interactions across e-commerce, journalism, social media, and emerging digital spaces (Bahtar & Muda, 2016; Santos, 2022; Wardle & Williams, 2010). In its early stages, UGC was primarily static, consisting of blog posts, product reviews, and forum discussions. However, with the emergence of Web 2.0, it has evolved into dynamic, multimedia-driven contributions, particularly on platforms such as Instagram, YouTube, and TikTok (M. Xu et al., 2022).

The relationship between UGC, social media, and business intelligence is best understood as a convergence of two domains, as illustrated in Figure 2.1. Social media platforms serve as the primary space for content creation and user interaction, enabling individuals to share opinions, experiences, and discussions (González-Bailón & Lelkes, 2023). Platforms like TikTok, Instagram, and X (Twitter) generate vast amounts of UGC that reflect user preferences, trends, and societal discourse. Meanwhile, BI focuses on analyzing and extracting insights from data, including structured business records and unstructured social media content (Grossmann & Rinderle-Ma, 2015). However, it is critical to note that while UGC serves as a bridge, not all user-generated content contains business intelligence. Much of the content shared on social media consists of personal interactions, non-commercial chatter, or "noise" that lacks strategic value for an organization. Therefore, the overlapping section in Figure 2.1 represents the specific subset of UGC that is relevant, actionable, and capable of being transformed into intelligence through data analytics and sentiment analysis.

By applying data analytics, sentiment analysis, and predictive modeling, BI helps organizations make informed strategic decisions based on emerging trends, consumer behavior, and market dynamics (Ravichandran et al., 2022). The overlapping section represents User-Generated Content (UGC), which serves as the bridge between Social Media and BI. Social media provides the platform for UGC creation, while BI processes and analyses UGC to derive actionable insights. This integration allows

businesses to track consumer sentiment, identify brand perceptions, and enhance decision-making through data-driven strategies.

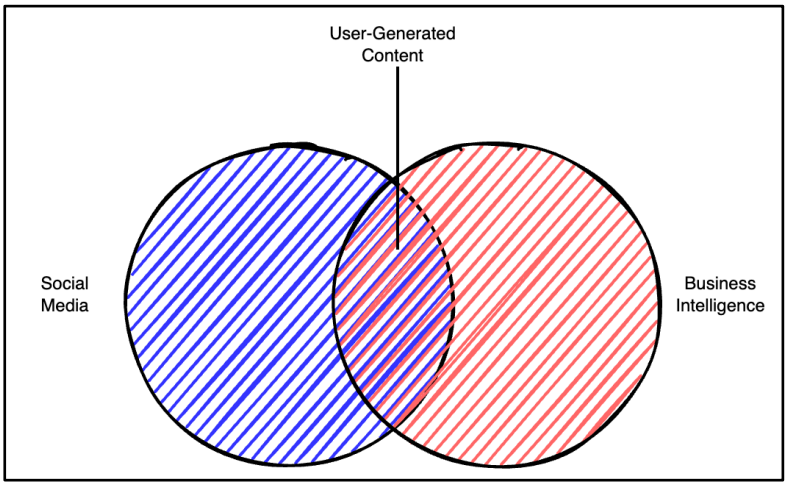


Figure 2.1 Venn Diagram Of The Role Of UGC As A Bridge Between Platforms And Intelligence

(Source: Authors illustration)

Recent advancements in artificial intelligence (AI) and blockchain technologies have transformed the nature of UGC, introducing both opportunities and challenges. The rise of AI-generated content (AIGC) has blurred the lines between authentic user contributions and automated content, raising concerns about manipulation and misinformation (Santos, 2022). Deepfake technologies and synthetic media have made it increasingly difficult to distinguish real content from fabricated material. In response, businesses and content platforms are exploring blockchain-based verification as a solution to authenticate digital assets and verify content ownership (M. Xu et al., 2022). Blockchain technology creates tamper-proof records, ensuring that UGC remains verifiable and resistant to manipulation. UGC has also expanded its influence in the hospitality and tourism sectors, where customer-generated reviews serve as real-time indicators of service quality. Businesses actively use UGC data to track customer satisfaction, detect service flaws, and optimize operational strategies (Maia et al., 2024). In this context, UGC functions not only as a marketing tool but also as a business intelligence resource, providing valuable insights into consumer behavior, brand perception, and emerging market trends.

The motivations for creating and sharing UGC vary across platforms, user demographics, and digital communities. Research has identified personal expression,

social validation, and peer influence as key drivers of content generation (Bahtar & Muda, 2016; Johnson et al., 2022). Many users engage with UGC to document personal experiences, express opinions, and participate in online communities. In e-commerce and business contexts, UGC serves a dual purpose, enabling consumers to share perspectives while providing businesses with insights into purchasing behaviours (Naab & Sehl, 2017). The concept of "produsage" describes how users continuously co-create and modify digital content, rather than passively consuming it (Naab & Sehl, 2017). This participatory culture is particularly evident in hospitality and tourism, where travellers generate photo diaries, travel blogs, and hotel reviews. Recent studies highlight that Gen Z travellers are among the most active creators of tourism-related UGC, often prioritizing emotional engagement over factual accuracy in their storytelling (Yamagishi et al., 2024). Unlike previous generations, younger consumers view digital storytelling as a social activity, using UGC to express identity, build credibility, and influence others.

The role of UGC in shaping consumer behavior is well-documented across multiple industries, particularly in e-commerce, news media, and travel. Social proof theory suggests that consumers rely on peer-generated content, such as reviews, testimonials, and influencer endorsements, when making purchasing decisions (Bahtar & Muda, 2016; Yew et al., 2018). Studies show that UGC is more trusted than corporate advertising, as authentic peer recommendations are often perceived as more persuasive and credible (M. Xu et al., 2022; Q. A. Xu et al., 2022). This phenomenon is also observed in journalism, where audience-generated content influences news credibility and information dissemination. However, news organizations maintain editorial control over which UGC is incorporated into reporting to ensure accuracy and ethical reporting standards (Naab & Sehl, 2017; Wardle & Williams, 2010). The hospitality industry provides a strong example of UGC's impact on consumer decision-making, as customer reviews on TripAdvisor, Booking.com, and Airbnb act as real-time service evaluations (Maia et al., 2024). Studies indicate that emotionally engaging UGC (EUGC) is more persuasive than factual UGC (FUGC) in influencing travel decisions. Consumers tend to prioritize personal narratives and peer testimonials over traditional marketing claims, reinforcing the idea that emotional storytelling plays a crucial role in digital consumer engagement (Yamagishi et al., 2024).

Additionally, AI-driven recommendation systems have intensified the influence of UGC by curating and amplifying content based on user preferences. These algorithms

ensure that consumers receive personalized content, further shaping buying decisions and online interactions (M. Xu et al., 2022). Despite its benefits, UGC raises concerns regarding credibility, authenticity, and misinformation. As deepfake technology and AI-generated content become more advanced, it is becoming increasingly difficult to differentiate between real and manipulated UGC (Santos, 2022). Consumers typically evaluate UGC based on source reputation, engagement levels, and platform verification mechanisms, but these factors alone do not eliminate the risk of misinformation (Bahtar & Muda, 2016; Shutsko, 2020). In journalism, strict editorial oversight is applied to verify audience-generated content, ensuring that inaccurate information does not spread through mainstream reporting (Wardle & Williams, 2010). However, in e-commerce and tourism, verification practices remain inconsistent, allowing for fabricated reviews, misleading influencer promotions, and manipulated consumer feedback (Maia et al., 2024; Yamagishi et al., 2024). To address these concerns, companies are exploring blockchain-backed verification systems and NFT-authenticated UGC, providing tamper-proof mechanisms to enhance content authenticity (M. Xu et al., 2022). Additionally, machine learning models are increasingly used in the hospitality sector to conduct sentiment analysis on UGC, allowing businesses to identify patterns of dissatisfaction and refine service strategies based on real-time consumer feedback (Bharadiya, 2023; Maia et al., 2024).

User-generated content exists at the intersection of social media, business intelligence, and consumer engagement, playing a key role in shaping brand perception, decision-making, and digital trust. It has transformed industries by providing real-time insights into consumer behavior, improving brand interactions, and serving as a valuable data source for business intelligence applications. However, the growing reliance on algorithmic amplification, AI-generated content, and decentralized platforms raises concerns about credibility, misinformation, and ethical governance. While AI-driven verification, sentiment analysis, and blockchain-backed authentication offer solutions, challenges such as content manipulation, privacy risks, and corporate influence over UGC moderation persist. Despite these complexities, UGC remains a powerful force in digital ecosystems, influencing consumer behavior, industry practices, and the evolving nature of data-driven business engagement.

2.4 Evolution of Social Media and Business Intelligence

The evolution of research in social media and business intelligence (BI) has been shaped by technological advancements, emerging digital platforms, and the growing reliance on user-generated content (UGC) for decision-making. As social media continues to influence communication and commerce, researchers have increasingly focused on its role in data analytics, sentiment analysis, and business intelligence applications, leading to an interdisciplinary approach that integrates insights from computer science, business studies, and information systems. To understand these shifts, bibliometric and text analysis methods provide a systematic way to examine publication trends, keyword evolution, and topic modeling results, helping to identify emerging research areas, track changes in dominant themes, and reveal underexplored topics. By merging insights from historical publication trends, research field contributions, and evolving research topics, this section presents a comprehensive analysis of how research in this field has transformed over time.

2.4.1 Growth of Social Media and Business Intelligence Research

The academic research on social media and business intelligence has grown significantly over the years, as illustrated in Figure 2.2. This increase reflects the rising importance of social media as a data source for business intelligence applications and the growing role of AI-driven analytics in decision-making.

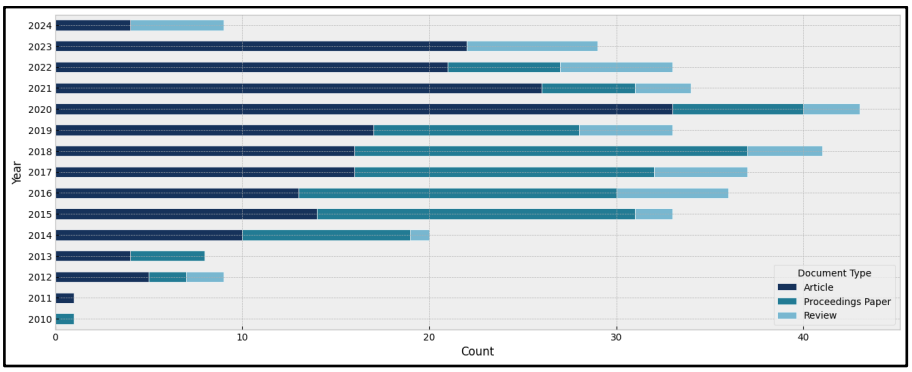


Figure 2.2 Documents Per Year

Among the different document types, journal articles dominate, making up the largest share of publications. The number of articles increased from one in 2011 to a

peak of 33 in 2020, showing a strong and sustained interest in the topic. This rise can be linked to technological advancements in AI, machine learning, and sentiment analysis, which have enabled businesses to analyze social media data more effectively. However, after 2020, there was a sharp decline, with only four articles published in 2024. This decline may be due to shifting research priorities, changes in funding allocation, or the emergence of new research areas. Despite this, the consistently high number of journal articles over the years confirms their role as a primary source for theoretical and empirical contributions in the field.

Proceedings papers, which are often presented at academic conferences, have shown fluctuations over time. The highest number of proceedings papers was recorded in 2018, with 21 papers, before declining in the following years. The reduction in proceedings papers after 2018 could be attributed to the COVID-19 pandemic, which significantly affected academic conferences and in-person research collaborations. As conferences are a major venue for presenting early-stage research, this decline suggests that researchers may have shifted towards alternative publication formats, such as journal articles or preprint servers. Review papers, although fewer in number, have maintained a consistent presence, with one to seven publications annually. Review papers play a crucial role in summarizing existing research, identifying trends, and highlighting gaps in the literature. The steady number of review papers over the years suggests that researchers continue to find value in synthesizing past studies to inform new research directions. Given the rapid evolution of AI and digital platforms, review studies remain essential for understanding the broader trends in social media and business intelligence research.

The impact of research in social media and business intelligence can also be assessed through citation trends, as shown in Figure 2.3. The number of citations reflects how often academic works are referenced in subsequent studies, indicating their influence in shaping the field.

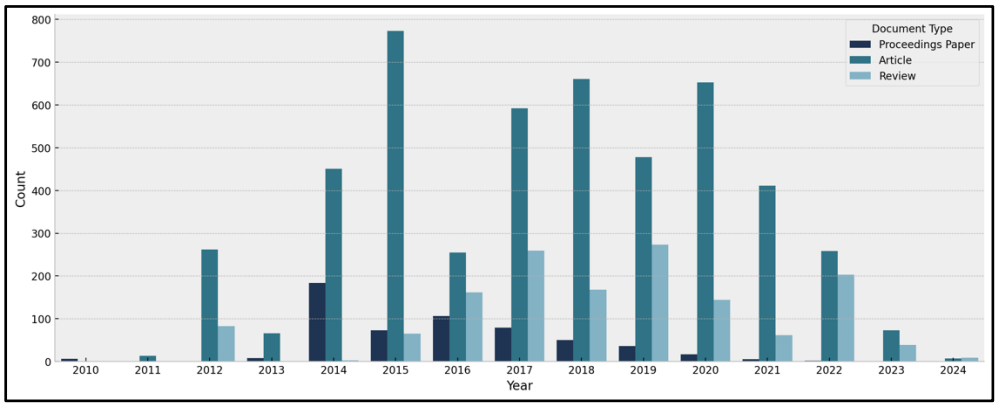


Figure 2.3 Citations Per Year

Journal articles have consistently received the highest number of citations, peaking at 773 in 2015. This peak suggests that key foundational studies were published around this time, laying the groundwork for future research. However, similar to publication trends, article citations declined sharply after 2020, with only seven citations recorded in 2024. This drop may indicate a shift in research focus, where newer studies are exploring different themes, or it could reflect a natural decline in citations for older publications as the field progresses. Review papers, in contrast, have shown a more stable citation pattern, peaking at 273 citations in 2019. This suggests that researchers continue to rely on review studies as reference materials, particularly when exploring new research areas. The stable citation trend of review papers also reflects their long-term relevance in providing comprehensive overviews of the field.

Proceedings papers have experienced a declining trend in citations, with the highest recorded citations occurring in 2014 (184 citations). This decline suggests that while proceedings papers contribute to early-stage discussions, they may have a shorter impact lifespan compared to journal articles. The reduced reliance on conference papers as primary sources of information indicates a preference for peer-reviewed journal articles as more credible and long-lasting references. The interdisciplinary nature of social media and business intelligence research is evident in the diverse range of contributing fields, as depicted in Figure 2.4. Research in this domain is driven by advancements in computer science, business studies, engineering, information systems, and information science.

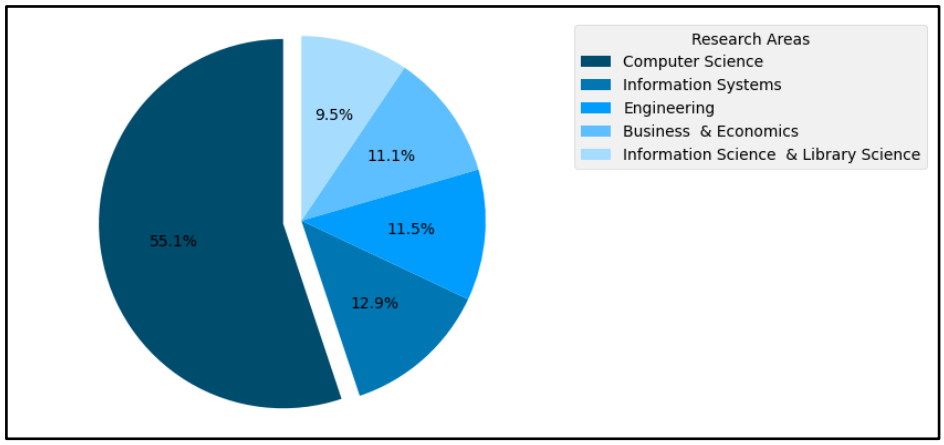


Figure 2.4Research Areas

- **Computer Science (55.1%)**: This field is the dominant contributor to research in social media and BI. It focuses on machine learning, big data analytics, artificial intelligence, and natural language processing (NLP) which are all essential technologies for analyzing social media data. The rapid development of computational tools has positioned computer science as the foundation of BI research, enabling businesses to process large-scale social media data for sentiment analysis and decision-making.
- **Business & Economics (12.9%)**: This area focuses on the practical applications of social media data in business intelligence, digital marketing, and market forecasting. Research in this field examines consumer behavior, competitive strategies, and social media-driven business models, helping businesses leverage data-driven insights for decision-making. The increasing role of AI-powered marketing strategies has further elevated the importance of this field in BI research.
- **Engineering (11.5%)**: The integration of social media data into system design and digital infrastructures has made engineering research relevant to BI. Studies in this area explore the development of scalable, high-performance computing solutions for handling large datasets and optimizing AI-driven decision-making systems.
- **Information Systems (11.1%)**: This field examines how **b**usiness intelligence systems are designed, implemented, and optimized. Research focuses on real-time data processing, database management, and cloud computing, enabling organizations to extract insights from social media platforms efficiently.

- **Information Science & Library Science (9.5%)**: Given the vast amount of unstructured data generated on social media, research in this field focuses on data management, retrieval, and structuring. The increasing need for frameworks that organize and store social media data efficiently has made this field an essential component of BI research.

The dominance of computer science and business & economics highlights the dual nature of social media and BI research, balancing technological innovation with practical business applications. Meanwhile, contributions from engineering, information systems, and information science reinforce the collaborative approach needed to develop scalable and effective BI systems. The changing research focus in social media and business intelligence can be observed through the Topics Over Time trends in Figure 2.5.

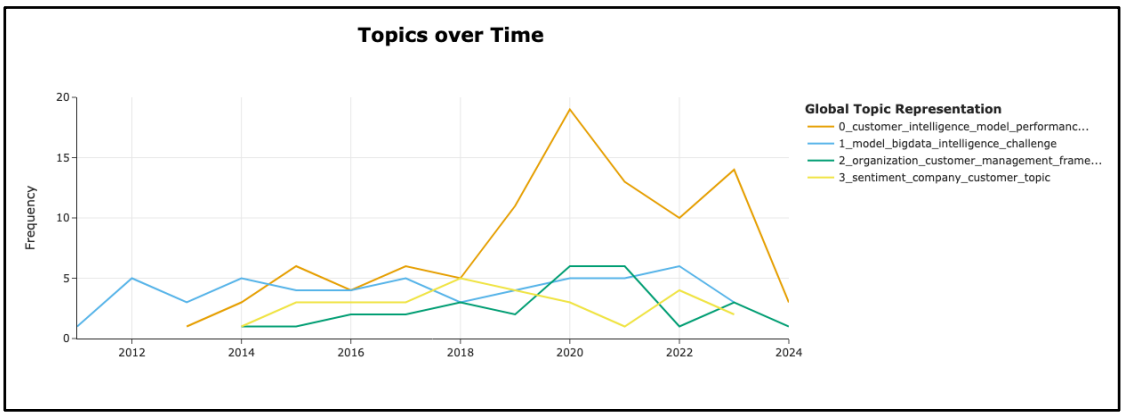


Figure 2.5 Topic Over Time

- **Customer Intelligence and Sentiment Analysis**: This topic has grown steadily from 2011 to 2019, with major peaks in 2019 and 2023. The 2019 peak aligns with the increased adoption of AI-driven sentiment analysis and big data tools, allowing businesses to analyze consumer emotions more effectively. The 2023 spike suggests further advancements in real-time AI-based sentiment tracking, which has become a key tool for businesses in consumer insights and competitive strategy development.
- **Big Data and Behavioural Insights**: This topic has been consistently relevant since 2011, peaking in 2012 and 2019-2021. The 2012 peak coincides with the early adoption of big data technologies, while the 2019-2021 surge corresponds

to behavioural changes during the COVID-19 pandemic. The pandemic led to shifts in consumer behavior, online engagement, and social media interaction patterns, which fuelled interest in behavioural analytics.

- **Customer Engagement and Innovation**: Gaining prominence in **2018**, this topic peaked in 2020, reflecting the surge in digital engagement strategies during the pandemic. Businesses relied heavily on social media marketing, AI-powered CRM systems, and automated customer support tools to maintain relationships with consumers.
- **Sentiment Analysis and Customer-Centric Models**: This topic has been stable since 2014, with notable peaks in 2018 and 2022. The 2018 peak reflects the growing importance of AI-powered sentiment analysis in business intelligence, while the 2022 peak aligns with the rise of advanced NLP models, chatbots, and AI-driven customer feedback systems.

These evolving research trends indicate an increasing focus on AI-driven insights, consumer analytics, and behavioural prediction models in social media and BI research. The growing reliance on real-time data processing and sentiment adaptation suggests that businesses and researchers will continue developing new methodologies for data-driven decision-making.

2.4.2 Connecting the Literature Review Findings

The evolution of research in social media and business intelligence has been driven by rapid advancements in artificial intelligence, big data analytics, and changing consumer behaviours (Praful Bharadiya, 2023; Shien et al., 2023). The findings reveal a clear shift toward AI-driven business intelligence, predictive analytics, and automated social media monitoring, while traditional studies focusing on engagement metrics and manual analysis have declined. This suggests that researchers are prioritizing technology-driven methodologies that enable businesses to analyze large-scale social media data more effectively. The analysis of research areas and topic evolution highlights several important trends. The increasing prominence of AI, sentiment analysis, and influencer marketing reflects the growing role of machine learning in extracting insights from social media interactions (Abdullah et al., 2023). At the same time, earlier business intelligence frameworks that relied on descriptive analytics are

becoming less relevant, as businesses move toward real-time, AI-powered decision-making. Thematic analysis using GPT-4o further refines these topics, demonstrating a shift toward research focused on adaptive consumer intelligence, algorithm-driven business strategies, and predictive analytics.

A strong relationship exists between research areas and dominant topics, indicating an interdisciplinary approach in this field. Computer science research primarily focuses on developing AI models for sentiment analysis, predictive analytics, and big data processing, while business and economics research applies these models to consumer behavior analysis, digital marketing, and strategic decision-making (Sarker, 2021; Shahadat Hosen et al., 2024). Additionally, interdisciplinary research is expanding, particularly in social network analysis, algorithmic trust, and AI ethics (Adanyin, 2024; Y. Zhang et al., 2021). The increasing focus on ethical and regulatory concerns in AI-driven social media analytics suggests that researchers are actively addressing risks associated with algorithmic bias and data privacy (Hancock et al., 2020; Saura et al., 2023). Despite these advancements, several research gaps remain. One of the most pressing gaps is real-time sentiment adaptation, as current AI models still struggle to adjust sentiment analysis dynamically in response to evolving consumer emotions (Rodríguez-Ibánez et al., 2023). AI bias in business intelligence is another critical issue, as algorithms trained on biased datasets can lead to misleading business insights and discriminatory outcomes (Camilleri, 2024). Additionally, privacy concerns related to large-scale social media data analysis continue to challenge researchers, particularly with increasing regulations like GDPR and CCPA limiting data accessibility (Di Minin et al., 2021; Saura et al., 2023). The GPT-4o summarization underscores these gaps, revealing the need for critical human validation in AI-driven research. While AI-generated insights are valuable for thematic analysis, they often lack contextual depth and may overlook nuances that require expert interpretation.

In conclusion, the research landscape in social media and business intelligence has undergone a significant transformation, with a strong shift toward AI-driven predictive analytics, real-time consumer insights, and automated business intelligence models. The most influential research areas include big data analytics, AI-powered consumer intelligence, and digital engagement strategies, reflecting the increasing reliance on technology-driven decision-making in business intelligence. As businesses and researchers continue to integrate AI into social media analytics, understanding

emerging trends and addressing key challenges will remain essential for future developments in this field.

2.5 Research Impact in Social Media and Business Intelligence

Table 2.2 presents the top five most influential articles in the social media and business intelligence field, ranked by metrics such as Times Cited, Usage-Count-Last-180-Days, and Usage-Count-Since-2013. These metrics provide insight into the academic and practical relevance of these works, showcasing the diverse methodologies and applications that have shaped the field.

Table 2.2
Top 5 High Impact Articles

Title	Times-Cited	Usage-Count-Last-180-days	Usage-Count-Since-2013
Exploring Trends and Patterns of Popularity Stage Evolution in Social Media	21	5	2248
Digital academic entrepreneurship: The potential of digital technologies in academic entrepreneurship	181	60	591
Business intelligence in online customer textual reviews: Understanding consumer engagement	136	8	406
Creating value from Social Big Data: Implications for Competitive Advantage	201	42	399
A novel social media competitive analytics framework with sentiment benchmarks	138	8	360

The first article, “Exploring Trends and Patterns of Popularity Stage Evolution in Social Media” (Kong et al., 2020), demonstrates the practical utility of predictive algorithms for analyzing social media trends. Although it has a relatively low Times

Cited count of 21, its exceptionally high Usage-Count-Since-2013 of 2248 highlights its continued relevance in guiding digital marketers and data scientists. The paper provides actionable insights for businesses to predict which content resonates best with their target audiences, enabling them to stay competitive in fast-evolving online environments. The second article, “Digital Academic Entrepreneurship: The Potential of Digital Technologies in Academic Entrepreneurship” (Rippa & Secundo, 2019), is one of the most cited, with 181 Times Cited and the highest Usage-Count-Last-180-Days at 60. This work focuses on the transformative role of digital technologies in academic entrepreneurship, especially relevant during the COVID-19 pandemic, when educational institutions were forced to adapt to online platforms. The paper offers insights for improving the effectiveness of digital learning environments, which is valuable for both academia and edtech businesses.

Another highly impactful research, “Business Intelligence in Online Customer Textual Reviews: Understanding Consumer Engagement” (X. Xu et al., 2017), has 136 Times Cited and a substantial Usage-Count-Since-2013 of 406. This article highlights the role of text analytics in interpreting consumer feedback, a critical area for businesses employing natural language processing (NLP) techniques. By leveraging such tools, organizations can improve their products and services based on customer preferences, demonstrating the practical value of integrating consumer insights into business strategies.

2.6 Comparing Social Media Platforms

Social media platforms vary significantly in content formats, audience interactions, and engagement dynamics, shaping their relevance in both academic research and business intelligence applications. Through bibliometric analysis, platform-related keywords reveal that studies have predominantly focused on Twitter and Facebook, while newer platforms like TikTok remain relatively underexplored despite their increasing influence. Comparing these platforms highlights their distinct characteristics, research potential, and strategic value for businesses.

Twitter is recognized for its text-centric format and real-time information flow, making it a primary choice for sentiment analysis, opinion mining, and social network analysis (Sanchez-Nunez et al., 2020; Tavoschi et al., 2020). Its open API access has allowed researchers to examine public discourse**,** crisis communication, and political

sentiment at scale. Businesses leverage Twitter's fast-paced interactions to track trends, respond to customer feedback, and analyze market sentiment in near real-time (Ravichandran et al., 2022; Uma Maheswari & Dhenakaran, 2023). Transitioning from a text-driven platform to a more visual and interactive space, Facebook provides a rich dataset combining text, images, videos, and community engagement. It is widely used for studies on user engagement, brand interactions, and targeted advertising. Facebook's structured demographic data makes it particularly valuable for consumer profiling, market segmentation, and business analytics, making it a key platform for digital marketing and customer relationship management (Kewsuwun & Kajornkasirat, 2022; Z. Xu et al., 2021).

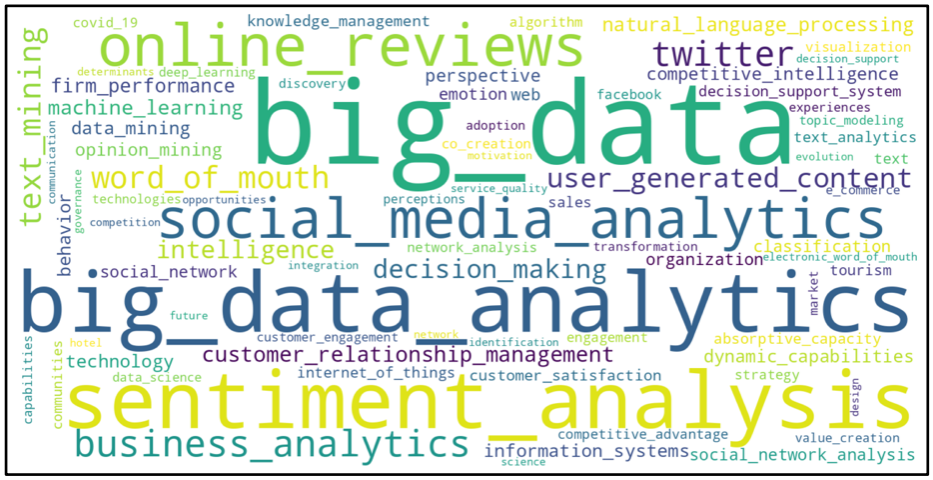


Figure 2.6 Word Cloud Of Author's Keywords

While Facebook and Twitter have long dominated academic research, newer platforms such as TikTok are reshaping digital engagement trends. Unlike its predecessors, TikTok is entirely video-driven, with a highly algorithmic recommendation system that personalizes content based on user behavior (Delmonaco et al., 2024; Harriger et al., 2023). Its emphasis on short-form videos, viral trends, and influencer-driven interactions presents new opportunities for studying consumer engagement, brand virality, and digital marketing strategies. However, its absence from dominant bibliometric keywords suggests that research on TikTok is still in its early stages, highlighting the need for further exploration. Beyond these three platforms, other social media networks, such as Instagram, YouTube, and LinkedIn, also contribute to business intelligence research (Feroz Khan & Vong, 2014; Thorgren et al., 2024). Instagram and YouTube focus on visual storytelling and influencer marketing,

while LinkedIn is valuable for professional networking and B2B engagement. Although these platforms hold significant analytical potential, TikTok's rapid growth and unique algorithm-driven engagement model make it a particularly relevant case for emerging research. Each platform presents unique opportunities and challenges for business intelligence and social media analytics. While Twitter and Facebook remain dominant in academic studies, the rise of TikTok and other emerging platforms signals a shifting digital landscape. As businesses and researchers seek more comprehensive consumer insights, integrating multiple platforms into analytical frameworks will be essential for developing adaptive and data-driven business intelligence strategies.

2.7 Techniques used in Social Media and Business Intelligence

Through bibliometric analysis, two key clusters of techniques emerge in social media and business intelligence (BI) research. The first cluster, Data Analytics & AI Techniques, represents computational methods used to process large-scale data, develop predictive models, and automate decision-making. The second cluster, Text & Sentiment Analysis in Social Contexts, focuses on understanding user-generated content (UGC), extracting opinions, and analyzing consumer sentiment. These techniques play a crucial role in helping businesses and researchers interpret social media data, track market trends, and enhance decision-making processes. By categorizing these methods into structured groups, the bibliometric findings provide a clearer understanding of the technological and analytical approaches shaping BI in the digital era.

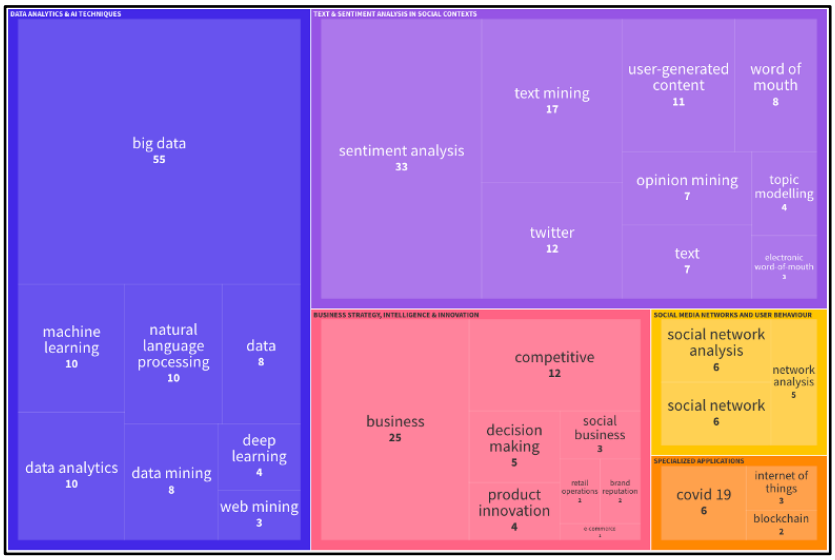


Figure 2.7 Clustered Tree Map Of Author's Keywords

2.7.1 Data Analytics and AI Techniques

The Data Analytics & AI Techniques cluster consists of computational methods that help businesses process vast amounts of social media data, extract meaningful insights, and improve decision-making. These techniques provide the technological foundation for business intelligence by enabling predictive modeling, automation, and large-scale data processing (Tariq et al., 2025). A central technique in this category is big data analytics, which allows organizations to handle structured and unstructured data from social media platforms. By integrating distributed computing, data warehousing, and predictive analytics, companies can analyze consumer behavior, track brand sentiment, and optimize marketing strategies in real time. Closely related to big data analytics, machine learning (ML) plays a critical role in automating data processing, detecting patterns, and developing predictive insights. Businesses use ML for customer segmentation, recommendation systems, and fraud detection, helping them adapt to changing market conditions (Talafidaryani et al., 2023).

Expanding on machine learning, deep learning (DL) enables more advanced data interpretation, particularly in tasks involving image recognition, natural language understanding, and automated decision-making. In social media analytics, deep learning is used for sentiment classification, chatbot interactions, and personalized content recommendations, allowing businesses to enhance customer engagement (Nanda & Kumar, 2021). Similarly, natural language processing (NLP) is a key technique for

analyzing textual data from tweets, product reviews, and online discussions. NLP-powered applications, such as chatbots, sentiment detection, and fake news identification, provide businesses with actionable insights from massive text-based datasets (Balakrishnan et al., 2021; Bauer & Clemm von Hohenberg, 2021). Another critical technique, data mining, uncovers hidden relationships within large datasets, enabling companies to extract actionable patterns for market research, competitive intelligence, and trend forecasting. Businesses apply data mining in consumer profiling and product demand prediction, helping them stay ahead of industry trends. Complementing data mining, web mining extracts insights from online sources, including social media interactions, customer reviews, and forum discussions. This technique allows companies to track brand reputation, detect emerging consumer trends, and optimize digital marketing strategies (Kanan et al., 2023; Lanza-Cruz et al., 2024).

Despite their advantages, these AI-driven techniques come with challenges. Bias in machine learning models can lead to inaccurate sentiment analysis, while privacy concerns in web mining raise ethical questions about data usage (Adanyin, 2024). Additionally, scalability issues in big data processing require businesses to invest in advanced computing infrastructure. Addressing these limitations is essential for ensuring that AI techniques are used responsibly and effectively in business intelligence applications.

2.7.2 Text and Sentiment Analysis in Social Context

The Text & Sentiment Analysis in Social Contexts cluster focuses on methods that help businesses interpret consumer opinions, analyze online interactions, and track social media sentiment trends. These techniques provide businesses with insight into public perception, brand positioning, and emerging market behaviours. A fundamental method in this category is sentiment analysis, which allows organizations to measure customer emotions and brand perception through social media data (Hartmann et al., 2023; Rodríguez-Ibánez et al., 2023). Businesses use sentiment analysis to track customer satisfaction, product reception, and public sentiment during marketing campaigns. Closely linked to sentiment analysis, opinion mining extracts subjective expressions from social media discussions, helping companies understand consumer attitudes, competitive positioning, and industry trends.

Topic modeling is another widely applied technique that identifies key themes within large datasets (Wang et al., 2024). Businesses use topic modeling to analyze customer complaints, trending discussions, and product feedback, allowing them to refine their services based on consumer needs. Social media analytics, which integrates multiple text analysis methods, enables companies to track engagement metrics, measure influencer impact, and optimize social media marketing efforts (Nanda & Kumar, 2021). A crucial tool in this category is text mining, which transforms unstructured data into structured insights. Businesses apply text mining to identify fake reviews, detect spam, and categorize customer inquiries, improving customer service and product innovation (He et al., 2013). Additionally, electronic word-of-mouth (eWOM), referring to the spread of consumer opinions on online platforms, has become a key factor influencing brand reputation and purchasing decisions. Companies analyze eWOM data to evaluate brand perception, measure campaign effectiveness, and predict consumer loyalty trends (Ye et al., 2011; Z. Zhang et al., 2010).

While these text-based analytics techniques offer valuable consumer insights, they also face challenges. Sarcasm, slang, and context ambiguity in sentiment analysis can lead to misclassification of opinions, affecting business decisions. Privacy concerns in eWOM and social media analytics raise ethical issues about data collection from online conversations. Furthermore, the evolving nature of language and internet slang makes it difficult for traditional text mining models to keep up with changing linguistic trends. To improve accuracy and reliability, businesses must combine AI-driven sentiment analysis with human validation methods (Aramburu et al., 2023).

The techniques used in social media and business intelligence can be categorized into data-driven AI techniques and text-based analysis methods. The Data Analytics & AI Techniques cluster focuses on handling large-scale datasets, developing predictive models, and automating insights, providing the computational power behind business intelligence. Meanwhile, the Text & Sentiment Analysis in Social Contexts cluster enables businesses to interpret consumer opinions, measure engagement, and analyze social interactions, ensuring companies remain competitive in digital markets. While these methods provide valuable insights, challenges such as bias, privacy concerns, and evolving data trends must be addressed to enhance their accuracy and effectiveness. The bibliometric findings confirm the significance of these methods in academic research and industry applications, highlighting their critical role in shaping modern business intelligence strategies.

2.8 Business Intelligence Strategies in Social Media

Social media has become a key source of business intelligence (BI), helping companies improve decision-making, marketing strategies, and customer engagement. Businesses use social media data to track competitors, analyze customer sentiment, and monitor industry trends (Amiri et al., 2023). By applying data analytics, artificial intelligence (AI), and machine learning (ML), companies can predict market changes, optimize marketing campaigns, and enhance their reputation management strategies. This section explores how businesses use BI strategies in social media, focusing on competitive intelligence, brand perception, marketing analytics, performance management, and AI-driven decision-making. Business intelligence in social media helps companies understand their competitors and market position. Competitive intelligence involves analyzing social media data to track industry trends and competitor strategies (Talafidaryani et al., 2023). Businesses monitor customer feedback, reviews, and online discussions to gain insights into consumer behavior. Predictive analytics allows companies to anticipate market demands and adjust their business strategies accordingly (Ali Hakami & Hosni Mahmoud, 2022). By using sentiment analysis and natural language processing (NLP), businesses can assess public opinions on their products and services compared to competitors. Brand perception and social influence also play a critical role in business intelligence. Consumers rely on social media for recommendations and reviews, making electronic word-of-mouth (eWOM) and influencer marketing powerful tools in shaping public opinion (Corallo et al., 2024). Companies use real-time sentiment analysis to monitor brand reputation and respond to negative feedback before it escalates. AI-powered tools help businesses track emerging trends and measure brand engagement across different platforms.

Marketing analytics and performance management allow companies to evaluate the effectiveness of their marketing campaigns. Businesses analyze customer demographics, engagement metrics, and purchasing behaviours to personalize marketing efforts. Machine learning models help identify high-value customers and predict purchasing patterns (Gesmundo et al., 2022). Additionally, companies measure key performance indicators (KPIs) such as customer sentiment scores, brand engagement levels, and advertising return on investment (ROI). By continuously tracking these metrics, businesses can refine their strategies and improve marketing efficiency. AI-driven business intelligence enhances decision-making by automating

data analysis and providing deeper insights (Praful Bharadiya, 2023). AI-powered business intelligence dashboards, chatbots, and predictive models allow businesses to respond to customer needs in real-time. Deep learning and NLP-based models help companies extract meaningful insights from large datasets, ensuring that decision-making is faster, more accurate, and based on real-time data. Business intelligence strategies in social media allow companies to monitor competitors, optimize marketing, enhance brand perception, and improve customer engagement. By leveraging competitive intelligence, sentiment analysis, marketing analytics, and AI-driven tools, businesses can make data-driven decisions that enhance their overall performance. As AI and data analytics continue to evolve, social media will remain a critical source of business intelligence, helping organizations stay competitive in the digital economy.

2.9 Specialised Applications of Social Media and Business Intelligence

Social media and business intelligence (BI) are not only used for marketing and competitive analysis but also for specialized applications in different industries as illustrated in the Figure 2.8 below. This figure provides a visual representation of these specialized applications through a keyword co-occurrence network. In this figure, the keywords represent the core themes identified in the literature, while the red numbers indicate the “co-occurrence” frequency which specifically, the number of academic papers within the dataset where these specific terms appeared together. For example, the frequency count of 5 for “decision making” and 3 for “competitive intelligence” highlights the relative weight of these topics in the current academic discourse. This mapping allows for an objective assessment of which specialized BI applications, such as e-commerce, crisis communication, and smart cities, are most prominently linked in existing research.

Emerging fields such as e-commerce, crisis communication, financial services, and smart cities rely on social media analytics to understand customer behavior, predict trends, and improve service delivery. This section explores the specialized applications of BI in social media, focusing on e-commerce and social commerce, crisis communication, digital transformation, and business process optimization. E-commerce and social commerce are among the fastest-growing applications of social media and BI (S. Singh & Sai Vijay, 2024). Businesses use social media analytics to track customer preferences, purchasing patterns, and online engagement. Consumer trust and brand

loyalty are key factors influencing online sales, and companies use sentiment analysis and opinion mining to assess customer satisfaction (Abdullah et al., 2023). AI-powered recommendation systems analyze social media interactions to provide personalized shopping experiences and targeted advertisements. Platforms like TikTok and Instagram have integrated social commerce features, allowing users to purchase products directly through social media.

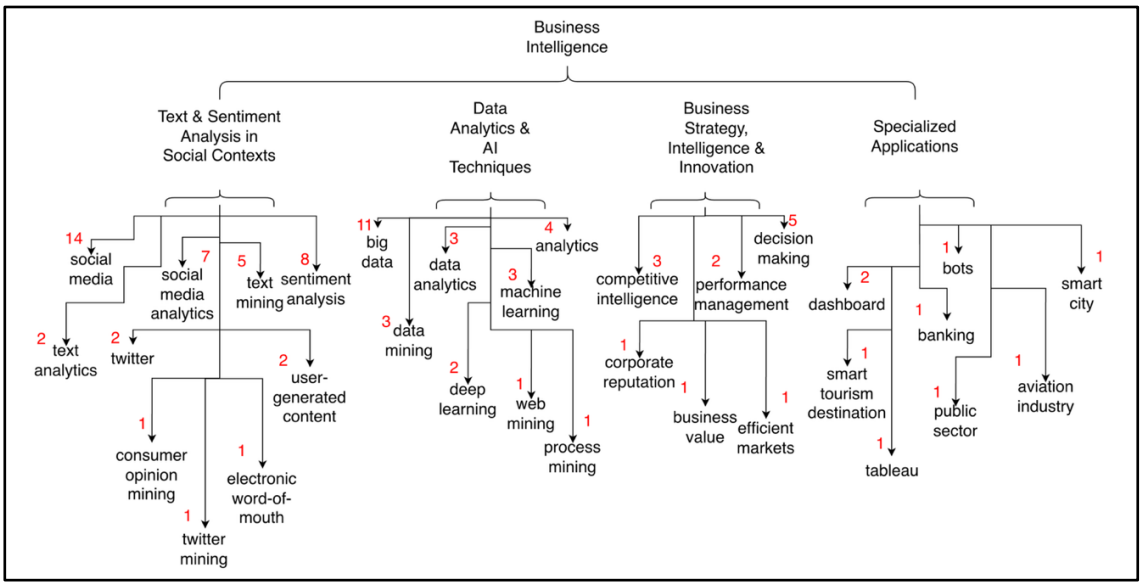


Figure 2.8 Keyword Co-Occurrences

Another important application of BI in social media is crisis communication and risk management. Organizations monitor social media for real-time updates on public sentiment, misinformation, and potential crises. During events such as pandemics, political instability, or product recalls, companies use AI-driven sentiment analysis to assess public reactions and adjust their communication strategies (Hartmann et al., 2023). Governments and public organizations also rely on social media analytics to manage public relations, address misinformation, and communicate emergency responses efficiently (Chan, 2024). Digital transformation and emerging business models are reshaping industries through social media and BI. Many businesses are shifting towards data-driven decision-making, integrating AI-powered BI tools into their operations. Social media provides insights into customer needs, product demand, and emerging market trends, helping businesses innovate and adapt. The financial sector, for example, uses social media sentiment analysis to predict stock market movements and assess investment risks. Similarly, the tourism and hospitality industries

leverage social media data to improve customer experiences and service offerings (T. Ali et al., 2022).

Finally, business process optimization and future trends highlight the role of BI in improving organizational efficiency, supply chain management, and resource allocation (Emmanuel Osamuyimen Eboigbe et al., 2023; Jahani et al., 2023). Smart cities use real-time social media data to manage traffic, monitor public services, and improve urban planning. Businesses also use predictive analytics to optimize logistics, reduce operational costs, and improve service delivery. As AI and machine learning continue to advance, automated BI systems will play an even bigger role in shaping industry-specific strategies. Social media and business intelligence are transforming various industries by providing real-time insights, improving decision-making, and enhancing operational efficiency. Applications in e-commerce, crisis communication, digital transformation, and business process optimization demonstrate the growing importance of social media analytics in business and governance. As technology evolves, businesses will continue to rely on AI-powered analytics to streamline operations, predict trends, and create more personalized experiences for consumers.

2.10 Research Framework

A framework serves as a conceptual model that structure ideas, methodologies, and analytical tools to support problem-solving, decision-making, and theory development across various academic and practical domains. In knowledge management and information technology, frameworks are essential for organizing complex data systems, guiding analytical workflows, and ensuring meaningful transformation of information into actionable knowledge (Baskarada & Koronios, 2013). Many disciplines rely on established frameworks to structure knowledge acquisition and decision-making processes. For example, the CRISP-DM model provides a framework for data mining workflows, while the Zachman Framework structures enterprise architecture to improve information system management. Similarly, the DIKW framework has become a fundamental model for understanding how raw data is processed into structured information, refined into knowledge and applied as wisdom (Ackoff, 1989; Rowley, 2007).

The Data, Information, Knowledge and Wisdom (DIKW) framework was first introduced by (Ackoff, 1989) and later refined by (Rowley, 2007). It describes a

hierarchical process where data is transformed into meaningful information, interpreted into knowledge and ultimately applied as wisdom. The DIKW framework has been widely adopted in business intelligence, AI-driven analytics, and industrial risk management, where structure knowledge processing is essential for decision-making (Lee et al., 2022; Werner et al., 2023). Originally, the DIKW framework assumed a linear, top-down structure, where each phase naturally progresses into the next. This framework has been applied in various fields including machine learning, healthcare decision-making, and cybersecurity, as a way to map information flow and improve decision quality (Cuomo et al., 2020). However, its rigid, hierarchical nature has been widely criticized, particularly in the era of AI-driven automation, generative AI, and digital twin modeling, where knowledge processing is no longer strictly sequential (Peters et al., 2024).

Primary criticism that DIKW framework faced is on its oversimplified structure and inability to accommodate modern AI-driven knowledge systems (Peters et al., 2024). Critics argue that knowledge is not merely accumulated information, and wisdom is not simple an advanced form of knowledge-rather, these phases involve interpretation, validation and social construction (Baskarada & Koronios, 2013). Furthermore, generative AI (Kim, 2024) and digital twin simulations (Grieves, 2024) challenge DIKW's assumption that data must be passively processed before being converted into actionable wisdom. To address these limitations, various scholars have proposed alternative interpretations of DIKW. These includes S-DIKW (Lo Duca & McDowell, 2024), which incorporates storytelling into knowledge retention; DIKWA (Lee et al., 2022), which adds an iterative "Action" phase for real-time decision feedback; and (Baskarada & Koronios, 2013) semiotic DIKW model, which emphasizes the interpretative nature of data and knowledge.

The following sections critically examine each phase of DIKW (Data, Information, Knowledge and Wisdom). By integrating insights from nine contemporary studies, this review explores how DIKW is evolving in response to AI-driven computational intelligence and digital knowledge ecosystems.

2.10.1 Data

At its core, refers to raw, unprocessed symbols, numbers, or observations that lack intrinsic meaning (Ackoff, 1989; Rowley, 2007). Traditionally, data is collected,

stored, and structured into information, but Baskarada & Koronios (2013) argue that data should not be seen as objective but as a semiotic construct, which is one that gains significance only through contextual interpretation and validation. Their semiotic DIKW framework proposes that data quality should not be assessed solely in terms of accuracy but must account for coherence, contextual relevance, and symbolic meaning. The rise of generative AI Kim (2024) has introduced new complexities to the definition of data, as AI now creates synthetic datasets that may contain fabrications, biases, or misinformation. Generative AI can produce seemingly accurate but entirely fabricated data points , which has raised significant concerns in scientific research, legal AI, and financing modeling. Thus, data integrity frameworks must evolve to account for AI-generated content validation.

Additionally, Digital Twin Systems (Grieves, 2024), has redefined data by treating it as an active component of real-time simulation models. In industrial applications, data continuously updates through AI-driven sensors, refining system efficiency and predicting failures. Unlike traditional static data repositories, Digital Twins create real-time, self-updating data streams, fundamentally changing how data is used in predictive analytics and industrial automation. Overall, in modern AI-driven environments, data is no longer a passive input but rather a dynamic, continuously generated and interpreted resource. Whether through semiotic analysis, AI-driven synthesis, or real-time simulation modeling, data today requires robust validation mechanisms to ensure accuracy, ethical application and contextual reliability (Baskarada & Koronios, 2013; Grieves, 2024; Kim, 2024).

2.10.2 Information

Information is formed when raw data is structured, categorised, and placed within a meaningful context (Ackoff, 1989; Rowley, 2007). While data exists as isolated observations, information provides answers to specific questions such who, what, where, and when, thereby enabling basic decision-making and pattern recognition. However, scholars argue that the transition from data to information is often ambiguous, as interpretation depends on the observer's cognitive model and contextual understanding (Baskarada & Koronios, 2013). A major challenge in modern information processing is the Data Rich, Information Poor (DRIP) problem, where organizations collect massive volumes of data but fail to extract meaningful insights

(Lee et al., 2022). This issue is particularly evident in industrial safety, cybersecurity, and AI-driven analytics, where excessive, unstructured data can obscure critical patterns rather that enhance decision-making.

Furthermore, Cuomo et al. (2020) introduce user-generated content (UGC) as a unique form of decentralized information structuring. In digital healthcare, for example, UGC allows patient reviews, online discussions, and medical forums to function as collective knowledge-sharing systems. Unlike traditional top-down knowledge repositories, UGC represents real-time, socially validated information flows, but also present challenges related to bias, misinformation, and source credibility. AI-generated information further complicates this landscape. Peters et al. (2024) highlights how AI models frequently generate dales yet plausible-sounding information, leading to issues of misinformation and AI bias. Meanwhile, Kim (2024) proposes the Generative AI-assisted Learning (GAIAL) model, which emphasizes human oversights in AI-generated information validation.

Thus, in contemporary AI-driven environments, the quality and authenticity of information depend on robust validation mechanisms that ensure contextual reliability, factual accuracy, and ethical application. This requires a multilayered approach, combining human expertise, algorithmic transparency, and social validation mechanisms (Cuomo et al., 2020; Kim, 2024; Lee et al., 2022).

2.10.3 Knowledge

Knowledge represents the meaningful synthesis of information that enables expert decision-making, predictive modeling, and problem solving (Rowley, 2007). Unlike information, which is structured data, knowledge integrates interpretation, expertise, and domain-specific reasoning to create actionable insights. The integration of AI into knowledge synthesis has significantly altered traditional frameworks. Kim (2024) introduces Data, Information, Knowledge, Intelligence and Wisdom (DIKIW) framework to accommodate AI-driven knowledge generation. AI can now analyse, synthesize, and generate knowledge autonomously, yet remains incapable of deep contextual reasoning, ethical discernment, or human-like judgment (Peters et al., 2024).

A contrasting approach comes from Grieves (2024), who explores knowledge synthesis in Digital Twins. In this framework, knowledge is continuously refined in real-time, as Digital Twins dynamically update sensor data, industrial analytics, and AI-

driven models to create and evolving predictive intelligence system. Despite advancements in AI-driven knowledge generation, Baskarada & Koronios (2013) stress that knowledge must still undergo social validation, ethical assessment, and expert review before being applied to critical decision-making. Thus, modern knowledge systems must balance AI automation with human oversight, ensuring that knowledge remains contextually relevant, ethically sound, and strategically useful in high stakes applications.

2.10.4 Wisdom

Wisdom, the highest level of DIKW, has traditionally been associated with strategic foresight, ethical reasoning, and human-centric decision-making (Ackoff, 1989; Rowley, 2007). Unlike knowledge, which analytical and explanatory, wisdom is prescriptive which guides long-term decision-making, moral judgement, and adaptive reasoning. However, Baskarada & Koronios (2013) argue that wisdom is not an endpoint but a dynamic, context-dependent construct which requires continuous validation through ethical frameworks, social norms, and experiential learning. Peters et al. (2024) strongly critique the assumption that AI can achieve wisdom, emphasizing that AI cannot engage in moral reasoning, self-awareness, or ethical judgment. While AI can optimize knowledge processing, automate decision-making, and provide statistical risk assessments, it remains incapable of comprehensive human values, emotions, or existential dilemmas. Kim (2024) introduces that GAIAL and GAIAT models, which reinforce the AI-assisted tasks and learning environments require human oversight to ensure responsible decision-making.

Additionally, wisdom in Digital Twins (Grieves, 2024) is not an inherent feature but a function of simulated knowledge refinement. Digital Twins can predict, test, and optimize industrial processes, but their decision-making logic is constrained by predefined models and statistical parameters. This means that Digital Twins, much like AI systems, require human supervision to ensure that simulations align with ethical considerations, safety standards, and long-term organizational goals. Thus, while AI can enhance decision-making processes, true wisdom remains uniquely human attribute, demanding more responsibility, strategic vision, and ethical reasoning that AI currently lack. The challenge lies in ensuring that AI-assisted decision-making aligns

with human values, preventing automations from overriding ethical constraints in critical domains such as governance, healthcare, environmental policy, and security.

This research demonstrates that DIKW is no longer a static hierarchy, but an evolving framework shaped by AI, Digital Twins, Generative AI, and semiotic analysis. The traditional assumption that wisdom is simply accumulated knowledge is no longer applicable in AI-driven knowledge systems, where automated decision-making, synthetic data generation, and predictive modeling challenge existing knowledge validation frameworks (Peters et al., 2024). As AI becomes more involved in data structuring, knowledge synthesis, and decision support, it is critical to ensure that automation is balanced with human ethical reasoning and strategic judgment. Scholars such as (Baskarada & Koronios, 2013; Grieves, 2024; Kim, 2024) emphasize that the future of DIKW must be adaptive, semiotic-aware, and capable of integrating AI-human collaboration. Moving forward, DIKW should be treated not as a rigid model but as a dynamic, context-driven framework that continuously evolves to accommodate new digital ecosystems, ethical considerations, and AI-human decision-making paradigms.

2.11 Research Methodology

Research methodology is the systematic process used to conduct a study, providing a structured approach to data collection, analysis and interpretation. It ensures that research findings are valid, reliable, and reproducible, making it a critical component of both academic and applied studies (Weyant, 2022). A well-defined methodology allows researchers to make informed decisions, ensuring that the chosen methods align with the research objectives and provide meaningful insights. In data-driven research, the choice of methodology is particularly important due to the complexity and scale of modern datasets. With the increasing availability of big data, machine learning, and artificial intelligence (AI)-driven analytics, traditional research methods are often insufficient to handle high-volume, unstructured, and continuously evolving data sources (Martinez-Plumed et al., 2021). This is especially true for research involving social media data, where new content is generated rapidly, and analysis requires flexibility to adapt to emerging patterns.

To address these challenges, data science methodologies have been developed to accommodate iterative, dynamic, and scalable approaches to data analysis. One such approach is Data Science Trajectories (DST), a methodology designed to handle the

non-linear and exploratory nature of modern data science projects. Unlike traditional structured methodologies, DST allows for continuous refinement of research goals, real-time data integration, and adaptive analysis. The flexibility makes it well-suited for research that involves diverse and evolving data sources, such as those generated from social media platforms. Figure 2.9 below shows the DST map containing the Exploratory, CRISP-DM, and Data Management activities.

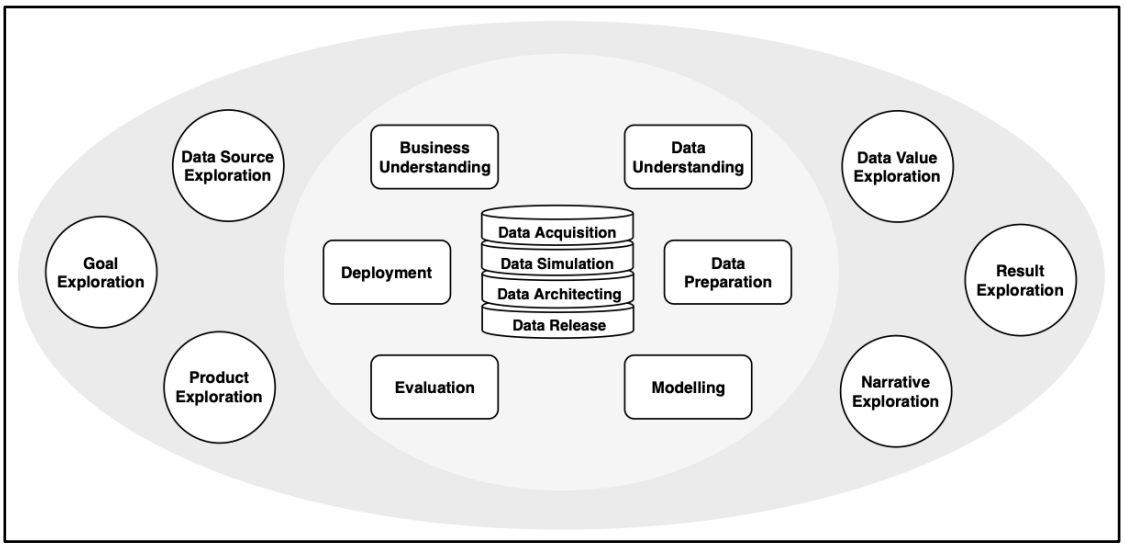


Figure 2.9 The DST Map Which The Outer Circle Is The Exploratory Activities, Inner Circle Is CRISP-DM Activities, And The Core Is The Data Management Activities

Source: (Martinez-Plumed et al., 2021)

The Data Science Trajectories (DST) model was developed to address the growing need for a more flexible approach in handling modern data science projects, particularly those that require continuous iteration and exploration. The CRISP-DM (Cross-Industry Standard Process for Data Mining) model, developed in the late 1990s, provided a linear and well-structured framework for data mining tasks (Martinez-Plumed et al., 2021). While effective for traditional projects, CRISP-DM followed a rigid, predefined set of steps, such as business understanding, data preparation, modeling, and evaluation, which limited its adaptability. As data science evolved, with more diverse data sources and more open-ended projects (Donoho, 2017), a new methodology was needed to address the challenges posed by the increasing complexity and exploratory nature of these project.

DST emerged as the solution to this challenge. It allows for non-linear processes, encouraging iteration and flexibility throughout the different stages of a data science

project (Martinez-Plumed et al., 2021). DST is particularly useful for handling modern data sources, such as those generated from social media, IoT devices, and other unconventional platforms, which require frequent revisiting of earlier phases to refine project objectives and methods based on emerging insights. DST offers significant improvements over CRISP-DM, particularly in handling more dynamic and exploratory data science projects. Table 2.3 below illustrates the key differences.

Table 2.3
Advantages Of DST Over CRISP-DM

Aspect	CRISP-DM	DST
Process	Linear and goal-driven with pre-defined stages	Iterative and flexible which allows for a non-linear movement between phases
Exploratory Capabilities	Assumes clear objectives from the start	Allows for exploration to refine goals based on data discoveries
Handling New Data Sources	Limited support for integrating new and unconventional data sources	Encourages discovering and incorporating new data sources such as social media
Data Management Activities	Focuses on data preparation and modelling	Incorporates data management tasks like acquisition, simulation, and architecting
Flexibility	Follows a fixed and structured process	Adapts to evolving project needs and enable for revisiting earlier stages

The DST model is designed with flexibility and exploration at its core. It acknowledges that data science projects are often iterative, with goals and methods evolving as new insights are discovered during the analysis (Martinez-Plumed et al., 2021). Unlike CRISP-DM, which follows a linear progression through predefined stages, DST allows for continuous movement between different phases of the project. Key components of the DST model include:

- Exploratory Activities: These are tasks that help refine the project's goals or discover new data sources. For example, goal exploration enables researchers to redefine the project’s objectives as the data reveals new insights. Data source exploration involves finding new and valuable data sources to enrich the analysis, which can significantly impact the project’s direction.
- Data Management Activities: These tasks are critical for ensuring that data is well-organized and prepared for analysis. Data acquisition involves collecting

relevant data, while data simulation can generate synthetic data to fill gaps. Data architecting focuses on integrating different datasets into a cohesive structure that allows for efficient analysis.

- Goal-Driven Activities: These include traditional tasks like data preparation, modeling, and evaluation, which contribute directly to the project's outputs. However, unlike CRISP-DM, these activities are not rigidly sequential and can be revisited as new data becomes available or as the project's objectives change.

The iterative nature of DST allows for a flexible approach to data science, where different phases can be revisited multiple times as necessary. This ensures that the project remains aligned with the data and insights as they emerge, making it highly adaptable for modern projects involving large and complex datasets (Martinez-Plumed et al., 2021). The DST diagram visually represents the iterative nature of the model, and the different types of activities involved. Each shape in the diagram has a specific meaning to show how these activities are interconnected and how the project can move fluidly between different tasks. This flexibility allows researchers to continuously refine their approach as new data and insights become available.

- Circles: Represent exploratory activities such as goal exploration and data source exploration. These activities are fluid and may be revisited multiple times throughout the project.
- Rounded rectangles: Indicate traditional goal-driven tasks like data preparation, modeling, and evaluation. These are essential steps that contribute directly to the success of the project.
- Cylinders: Symbolize data management activities such as data acquisition, data simulation, and data release. These tasks ensure that the necessary data is available, clean, and properly structured for analysis.

According to Martinez-Plumed et al. (2021), the DST chart is a directed graph where each activity occurs once, and transitions between activities are represented as directed arrows. Each transition between activities is numbered sequentially, ensuring no infinite loops, which preserves the linear progression of tasks. For example, a DST might begin with Goal Exploration (transition 0) and proceed to Business Understanding (transition 1), where parallel paths may diverge for distinct tasks like

data understanding and data exploration. The model allows revisiting activities without unnecessary loops; for instance, a trajectory might move from Modelling to Evaluation and return to Business Understanding for refinement, yet the transitions would only be repeated as necessary without creating an unmanageable process loop.

A practical case might involve exploring customer data for a business intelligence project. The project starts with goal setting (Goal Exploration), followed by Business Understanding (exploring customer behaviour insights), and transitions into Data Preparation. After the modelling phase, the process may revisit business understanding to fine-tune the model based on newly discovered insights, illustrating DST's flexibility in real-world use cases. This structure emphasizes that while data science requires iteration, DST charts ensure progression by limiting loops and enabling focused transitions between stages. This method is particularly advantageous compared to CRISP-DM's linear structure as it allows for more fluid and adaptable processes suited for complex, evolving data science workflows. Figure 3.1 below shows the DST map containing the Exploratory, CRISP-DM, and Data Management activities.

An example of the DST model in action can be found in a location-based tourism recommendation system which its methodology is illustrated in the figure 2.10. In this case, the project begins with goal exploration, where the team defines the type of personalized recommendations to offer users based on their location. In the data source exploration phase, relevant data is gathered, such as users' location histories and reviews of tourist spots. Next, data preparation involves structuring the data into a usable format for analysis, followed by modeling, which develops the recommendation engine. Lastly, product exploration focuses on delivering these recommendations, such as through a mobile app or website interface. This use case highlights the flexibility of the DST model, allowing the project team to adapt their objectives and methods as new data and insights emerge. The iterative nature of DST ensures that the final product is highly tailored to user needs and based on continuously refined data.

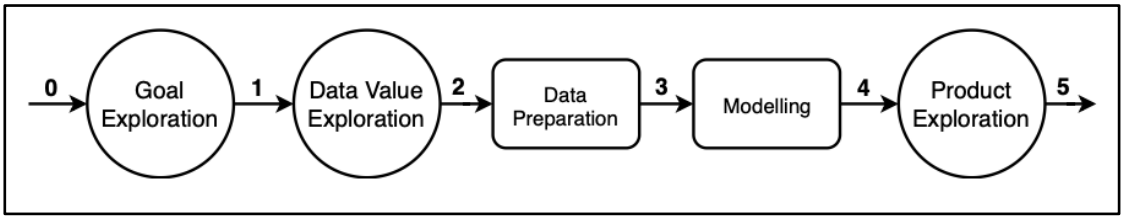


Figure 2.10 The Methodology Uses For Tourism Recommendation System

Source: (Martinez-Plumed et al., 2021)

Based on the Data Science Trajectories (DST) model (Martinez-Plumed et al., 2021) , the Social Media Business Intelligence Methodology was crafted specifically for this research, which focuses on the analysis of user-generated content (UGC) from social media platforms like TikTok. The iterative and flexible nature of DST makes it ideal for handling dynamic data sources like TikTok and allows for ongoing refinement of goals, data acquisition, and analysis methods as the project progresses. Below is the explanation of each step in the social media business intelligence methodology, as illustrated in the figure 2.11 below.

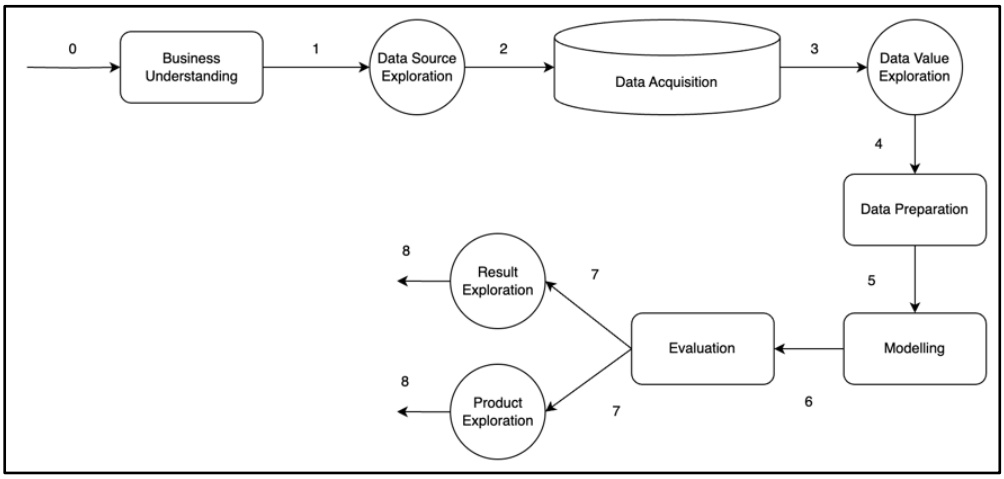


Figure 2.11 The SMBIM Adapted From DST Model

The social media business intelligence methodology begins with Business Understanding (transition: 0th), which involves a deep examination of the business problem. This stage defines the research's core objectives in the context of business intelligence, focusing on how businesses can leverage insights from TikTok's UGC to develop marketing strategies. In this research, the aim is to understand the landscape of social media and business intelligence, and the influence of consumer behaviour and key topics of discussions on brand perception, sales, and consumer engagement.

Once the business objectives are clear, the process moves to Data Source Exploration (transition: 1st), where potential data sources are identified and explored. TikTok, being a relatively underexplored platform in academic research, is chosen as the primary source of UGC for this research. The goal is to gather valuable data such as video metrics (likes, comments, views), influencer profiles, and consumer comments to

form the foundation of the analysis. This step emphasizes the importance of selecting relevant and valuable data sources that can provide insights into the influencer-consumer relationship on TikTok. Following data exploration, the process continues to Data Acquisition (transition: 2nd), where in this research the data is collected using tools such as Kalodata and Apify. Kalodata is used to gather revenue data for TikTok influencers, while Apify is employed to extract video metadata and user comments. These tools enable the collection of large volumes of structured data, which is critical for analysing both influencer activities and consumer responses.

After acquiring the necessary data, the next step is Data Value Exploration (transition: 3rd), where the data metrics are assessed to figure out what value might be extracted from the data. At this stage, initial data description analysis is conducted to gain the foundation understanding of the dataset, identify potential patterns, and determine how well the data aligns with the research objectives. In this research, special attention is given to metrics such as revenue generation, influencer engagement, and consumer feedback. Once the data has been explored, the methodology proceeds to Data Preparation (transition: 4th). In this phase, the data is cleaned, transformed, and pre-processed to ensure it is suitable for analysis. For instance, comment text from TikTok users may need to be standardized, with special attention given to slang, abbreviations, and misspellings in the local language, Malay. In addition, variables such as views, likes, shares, and comments are normalized to facilitate comparative analysis across influencers. Next, the data is ready for Modeling (transition: 5th). This phase involves applying machine learning and analytical techniques to derive insights from the data. In this research, modeling includes descriptive analytics to provide statistics overview of the data, the application of sentiment analysis to understand consumer opinions and topic modeling to identify key themes in consumer discussions related to beauty products.

Once the models have been developed, the process proceeds to Evaluation (transition: 6th), where the models' performance is assessed. This step is critical to ensure that the insights generated are accurate and actionable. In this research, the accuracy of sentiment analysis is measured based on the comparative analysis between GPT-4o model and specialized sentiment analysis model. Meanwhile, for topic modelling, coherence score is used to evaluate whether the topics are semantically categorised. The final steps in the methodology involve Result Exploration and Product Exploration (transition: 9th). In Result Exploration, the focus is on interpreting the

results and deriving actionable insights that can be applied to business strategies. For instance, insights from TikTok UGC can guide businesses in refining their marketing strategies or how consumer sentiment shifts around different products. These actionable recommendations help brands optimize their campaigns and make data-driven decisions for improved engagement and sales. In contrast, Product Exploration deals with developing dashboards or visualizations that can effectively communicate these insights to decision-makers. This stage involves designing visual representations of the data, such as interactive charts or reports, that allow businesses to monitor key performance indicators (KPIs) and track the success of their strategies. These tools enable stakeholders to explore the data on their own, gaining real-time access to the information they need to make informed decisions. By visualizing the impact of influencer campaigns and customer sentiment, businesses can better tailor their strategies to achieve their goals.

The SMBIM is guided by the DST model (Martinez-Plumed et al., 2021), allows for a flexible and iterative approach to data science, ensuring that the research objectives can be continuously refined and adapted as new data becomes available. The integration of exploratory and goal-driven activities enables a comprehensive analysis of TikTok's UGC and its impact. Table 2.4 below shows the supporting literature and description for each phases applied in this methodology.

Table 2.4
Description For Each Phases

Methodology Component	Phases	Tasks	Literature and Description	Transition
CRISP-DM	Business Understanding	- Identify the role of UGC on social media. - Define how TikTok will be used for UGC analysis.	The impact of data-driven decision-making on businesses is evident in how they can use TikTok data to gain insights into consumer behaviour and brand perception, as highlighted by Sarker, 2021 The role of user-generated content (UGC) is also crucial, with Kaplan & Haenlein, 2010b providing a foundational view on how user interactions can inform and enhance marketing strategies. Additionally, the importance of aligning project goals with business needs is discussed by de Mast & Lokkerbol, 2024 , whose framework proves useful in this context.	Transition 0: Sets core objectives.
DST	Data Source Exploration	- Explore social media platforms. - Assess TikTok's relevance as a source for business data.	Big data and knowledge discovery have proven valuable in large-scale production systems, as discussed by Abasova et al., 2021 which highlights the need to explore new and relevant data from social media platforms like TikTok. The contemporary analysis of TikTok's rise presented by (Li, 2022) underscores the platform's relevance in modern business intelligence. This perspective aligns well with the data source exploration phase, offering both academic and practical justification for selecting TikTok as the focal platform for this research.	Transition 2: Identifies suitable data sources for analysis.

DST	Data Acquisition	- Collect data from Kalodata and Apify. - Retrieve TikTok metadata and comments.	Methods for efficiently extracting data from social media platforms, as described by (Jin et al., 2015), are essential when gathering TikTok metadata and comments through web scraper tool such as Apify. Their work also highlights best practices in large-scale data collection and processing, which is particularly relevant when pulling data from multiple sources like Kalodata and Apify.	Transition 3: Data gathering phase, leveraging tools for structured data acquisition.
DST	Data Value Exploration	- Analyse dataset features. - Define insights from TikTok variables (e.g., comments, followers).	(Almeida et al., 2021) explore how social media data, particularly from platforms like TikTok, can yield insights through feature analysis such as comments, likes, and engagement metrics. (Jin et al., 2015) discuss knowledge discovery and emphasize the importance of extracting value from raw data, which fits with your task of analysing TikTok data features and translating them into meaningful business insights. (Zeng et al., 2010)'s foundational work on feature extraction further supports this phase.	Transition 4: Determines the data's utility and alignment with research goals.
CRISP-DM	Data Preparation	- Clean and transform data. - Handle missing or noisy data. - Prepare data for modeling.	(Schröer et al., 2021) provide a systematic review of the CRISP-DM process, specifically addressing the data preparation stage, making it directly relevant to tasks around cleaning and transforming data. (Rahm & Hong, 2000) discuss more recent challenges in handling unstructured data, such as spelling inconsistencies and missing data, common in social media datasets like TikTok comments. (Skarpathiotaki & Psannis, 2022) also cover best practices in text data preparation, which is crucial when preparing comments for analysis.	Transition 5: Data pre-processing to ready it for modeling and analysis.

CRISP-DM	Modelling	- Conduct Descriptive Analytics to understand data distribution and summary statistics. - Perform Sentiment Analysis using GPT and a specialized algorithm for Malay sentiment analysis. - Apply Topic Modeling using BERTopic to extract key themes from TikTok comments. - Analyse correlations and trends to assess relationships between variables.	**Descriptive Analytics:** Understanding the distribution of data through summary statistics and correlations is essential to any data analysis project. (Fan & Gordon, 2014) discuss how descriptive analytics forms the foundation for more advanced analytics, offering an overview of the data landscape, which is crucial for assessing patterns in social media metrics like views, likes, and comments on TikTok. (Abasova et al., 2021) further explain how descriptive analytics plays a vital role in production and decision-making systems, reinforcing its importance in this research. **Sentiment Analysis:** Analysing user sentiment in TikTok comments provides insights into public opinion on products or influencers. (Kaplan & Haenlein, 2010) discuss how sentiment analysis of UGC can offer businesses valuable feedback. More specifically, (Almeida et al., 2021) explore how modern sentiment analysis models can be combined with other techniques like topic modeling to generate deeper insights. This research make use of GPT for general sentiment analysis and a specialized algorithm for Malay sentiment analysis provides a dual-perspective approach, which is valuable in understanding sentiment across different linguistic contexts. **Topic Modeling:** (Blei et al., 2003) introduced Latent Dirichlet Allocation (LDA), a foundational topic modeling technique, but more recent advancements, like (Grootendorst, 2022)'s BERTopic, provide more nuanced	Transition 6: Analytical phase to extract insights and trends from the data.

			modeling approaches, particularly suited for short-form content like TikTok comments. (Skarpathiotaki & Psannis, 2022) also discuss cross-industry text analytics, providing justification for applying topic modeling to analyse social media content and extract meaningful topics related to the beauty and personal care industry. Using BERTopic, you can generate clusters of discussion around themes, such as product reviews, trends, or influencer promotions. Integrating these models allows you to assess not just what users are discussing but also the underlying sentiments and themes driving these conversations. (Graser et al., 2024) further highlight how combining multiple models (descriptive, sentiment, and topic) leads to richer, multidimensional insights, making the data more actionable for business purposes.	
CRISP-DM	Evaluation	- Validate the accuracy and reliability of the sentiment analysis models (GPT and specialized Malay algorithm). - Measure the coherence and diversity of topic models using standard evaluation metrics like coherence scores.	**Sentiment Analysis Evaluation:** Validating the accuracy and performance of sentiment models is critical for ensuring reliable insights. (Röder et al., 2015) discuss the importance of coherence measures in ensuring that models accurately reflect human judgment, a consideration that applies to both topic and sentiment models. For sentiment analysis, a comparison between GPT (a general-purpose model) and a specialized Malay sentiment model ensures that the nuances of the language are captured correctly. This dual approach mirrors the recommendations of (Syed & Spruit, 2017), who highlight the value of evaluating models against different linguistic and contextual baselines to ensure they are robust across varying scenarios. Sentiment models can be	Transition 7: Assesses the model's effectiveness and reliability.

		- Compare sentiment analysis results across models to assess consistency and accuracy. - Analyse the stability of topics over time or across different user segments.	evaluated through metrics such as accuracy, precision, recall, and F1-score, ensuring that both general and specific sentiments are captured effectively. **Topic Modeling Evaluation:** Evaluating topic models relies heavily on coherence and diversity scores to assess the quality of topics generated. (Röder et al., 2015) introduced the UMass coherence score, which has become a standard for evaluating topic models by measuring the semantic similarity of words within topics. Coherence scores ensure that the topics generated are meaningful and interpretable. In addition to coherence, (Syed & Spruit, 2017) propose combining coherence with diversity metrics to ensure that the model captures a wide range of themes from the dataset, preventing redundancy or overfitting. Furthermore, (Tripathi et al., 2021) highlight the need to evaluate the stability of topics over time or across different user segments, which is particularly relevant when analysing TikTok data, as trends evolve rapidly. Ensuring that topics remain consistent across different timeframes or demographic groups will enhance the reliability of your insights. Using these evaluation techniques allows you to ensure that your topic models provide robust, actionable themes that businesses can rely on for decision-making.	
DST	Result Exploration	- Interpret and analyse the model output.	(Almeida et al., 2021) expand on the interpretability of model outputs by linking topic models with sentiment analysis, helping generate richer insights	Transition 8: Provides insights to

		- Explore key insights from the topic and sentiment models.	from social media data. This aligns with your tasks of interpreting TikTok data and combining the results of topic and sentiment models for business purposes. (Sievert & Shirley, 2014)'s development of LDAvis also supports this phase by providing a way to visualize and explore topic model results interactively, adding value to your result exploration phase.	improve business outcomes based on research findings.
DST	Product Exploration	- Apply insights to business decision-making. - Explore how TikTok insights could impact marketing.	(Fan & Gordon, 2014) discuss how social media insights can be directly applied to business strategies, providing a solid foundation for your exploration of TikTok insights in business decision-making. (Peng et al., 2023) takes this further by examining how TikTok data specifically influences brand positioning and product development. The rapid feedback cycle on TikTok makes it a valuable tool for businesses, especially in industries like beauty and personal care where trends evolve quickly.	Transition 8: Strategic application of research findings in a business context.

2.12 Framework + Methodology

A well-structured research approach requires both a conceptual framework to establish theoretical underpinnings and a practical methodology to operationalize data collection, analysis, and interpretation. The Data, Information, Knowledge, and Wisdom (DIKW) framework serves as the conceptual foundation for this research, providing a structured representation of how raw data transform into meaningful insights and actionable wisdom. While DIKW outlines a theoretical approach to knowledge creation, it does not prescribe specific methodological steps to process and analyse data. To address this gap, the Social Media Business Intelligence Methodology (SMBIM) is employed as the practical methodology, allowing for the systematic implementation of DIKW principles in the context of TikTok user-generated content (UGC) analysis.

Integrating DIKW and SMBIM ensure that the research maintains both conceptual depth and methodological rigor. DIKW provides a structured way to categorise and interpret data within an information-processing hierarchy, while SMBIM ensures a practical, iterative, and goal-oriented approach to analysing social media data. By aligning each phase of DIKW with corresponding SMBIM processes, this integration facilitates a structured transition from data collection to business intelligence insights. The following sections detail how DIKW and SMBIM are interlinked, ensuring that raw, unstructured social media data evolves into actionable business intelligence in a structured manner. Figure 2.12 below show the how this research mapped the conceptual framework to the practical methodology.

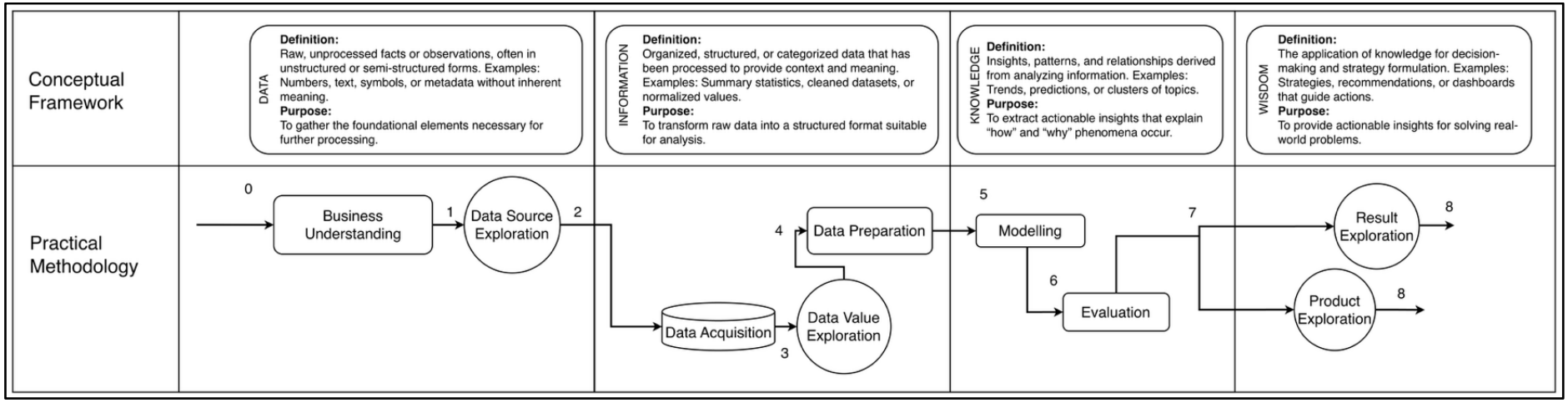


Figure 2.12 Map Conceptual Framework To The Practical Methodology

2.12.1 Conceptual and Practical Alignment of DIKW and SMBIM

The integration of DIKW and SMBIM is based on the premise that data, when processed systematically, can lead to knowledge that informs decision-making. DIKW provides a hierarchical model of data transformation, while SMBIM provides the steps and tools necessary to execute this transformation in practice. The following subsections outline how each phase of DIKW is mapped to specific processes in SMBIM, ensuring a seamless flow from data collection to wisdom-driven insights.

In the DIKW model, the data phase represents the raw, unstructured, and unprocessed data collected from various sources (Peters et al., 2024). Data lacks inherent meaning and must be organized and structured before it becomes useful. In the context of this research, TikTok UGC serves as the primary raw data, encompassing video metadata, engagement metrics (likes, shares, comments), influencer profiles, and consumer interactions. At this stage, data is collected in its raw form without contextual interpretation, making it essential to define research objectives and refine data collection methods. Within SMBIM, this phase is operationalised through three key processes which is Business Understanding, Data Source Exploration, and Data Acquisition. The Business Understanding phase establishes the core research questions and determines which aspects of TikTok engagement and influencer activities contribute to business intelligence. This includes questions such as how influencer engagement metrics correlate with revenue generation and how consumer sentiment influences brand perception. Once business objectives are defined, that Data Source Exploration phase identifies where relevant data can be obtained, ensuring that datasets align with the research's goals. Data Acquisition follows, using tools like Kalodata and Apify to extract structured and unstructured data from TikTok, including engagement trends, sentiment, and audience interactions. By structuring the data collection process through SMBIM, this phase ensures that high-quality, relevant, and comprehensive datasets are collected to support subsequent analysis.

The Information phase in the DIKW model represents structured and categorized data that provides meaning. Information is derived when raw data is cleaned, labelled, and formatted in a way that allow for interpretation. This phase is crucial in ensuring that irrelevant, duplicate, or low-quality data is removed, leaving behind datasets that can be analyzed for meaningful patterns (Peters et al., 2024; Rowley, 2007). Within SMBIM, two processes facilitate the transition from Data to

Information, which are Data Value Exploration, and Data Preparation. Data Value Exploration ensures that only meaningful and relevant data is retained by evaluating which metrics contribute to influencer engagement insights and brand impact. This involves filtering out irrelevant comments, eliminating noise in engagement metrics, and identifying key behavioural trends within TikTok UGC. Data Preparation follows, refining the dataset further by applying cleaning techniques such as handling missing values, normalizing engagement rates, and preprocessing text-based comments. Special attention is given to Malay UGC, ensuring that slang abbreviations, and informal expressions are standardised and interpreted accurately. By aligning information processing with SMBIM's structured data refinement methods, this phase ensures that raw social media content is transformed into structured, meaningful datasets that can be subjected to advanced analytics techniques in the next stage.

The knowledge phase in DIKW represents the generation of insights from processed information (Peters et al., 2024). At this stage, data is not just structured but is analysed to extract meaningful relationship, trends, and patterns. Knowledge answers "how" and "why" questions by revealing the impact of influencer engagement, sentiment analysis trends, and audience behaviours on TikTok (Rowley, 2007). Within SMBIM the transition from Information to Knowledge is operationalised through Modelling and Evaluation. The Modeling stage applies machine learning techniques such as Sentiment Analysis and Topic Modelling to categorise and analyse TikTok comments. Engagement trend analysis further examines patterns in likes, shares, and consumer sentiment across different influencers and content types. Once models are developed, the Evaluation stage ensures accuracy and reliability, using metrics such as sentiment accuracy comparisons and coherence scores for topic modelling. By leveraging SMBIM's structured modeling process, this phase ensures that analysed data produces validated knowledge that explains influencer engagement and consumer sentiment trends.

Wisdom represents the strategic application of knowledge to real-world decision-making. In DIKW, wisdom is the final phase where knowledge is translated into actionable insights that inform business strategies and decision-making frameworks (Peters et al., 2024). In SMBIM, this phase is realized through Result Exploration and Product Exploration. Result Exploration involves the interpretation of research findings to derive actionable recommendations for brands and influencers. It translates insights into business strategies, influencer marketing recommendations, and

engagement optimization techniques. Product Exploration then visualizes and communicates findings, developing business intelligence dashboards, interactive data reports, and key performance indicators (KPIs) that allow brands to track influencer impact and consumer sentiment in real time. By integrating DIKW's conceptual understanding of wisdom with SMBIM's structured decision-making framework, this phase ensures that research findings are transformed into meaningful actionable business intelligence applications.

2.13 Conclusion

This chapter has systematically examined the evolution of social media and business intelligence research, highlighting key methodologies, trends, and applications. The review provided a structured analysis of existing literature, identifying how social media platforms have emerged as a valuable data source for business intelligence, along with the diverse analytical approaches used to extract insights. Despite significant advancements, several critical research gaps remain. First, while social media has been extensively studied in business intelligence, most existing studies focus on established platforms such as Twitter and Facebook. Research on emerging platforms like TikTok remains relatively underexplored, particularly in the context of consumer engagement and business strategies. This gap presents an opportunity to examine TikTok as a unique data source and assess its role in shaping digital marketing and e-commerce strategies.

Second, current literature primarily employs traditional machine learning and statistical approaches for analyzing social media data. Although deep learning and transformer-based models such as GPT and BERT have gained traction, their application in social media business intelligence, particularly in understanding nuanced consumer sentiment and engagement patterns, is still developing. More research is needed to evaluate the effectiveness of these advanced models in extracting actionable insights. Third, while existing studies discuss various business intelligence strategies, there is limited research on integrating real-time social media analytics with decision-making frameworks. The dynamic nature of social media data requires adaptive models that can capture emerging trends and shifts in consumer behavior. Lastly, there is a lack of standardized frameworks for evaluating the impact of social media-driven business intelligence. While studies have explored various metrics such as engagement rates and

sentiment scores, a comprehensive framework that links these metrics to business performance indicators remains absent. Addressing this gap would provide a more structured approach to assessing the effectiveness of social media strategies in driving business outcomes.

In summary, this chapter has critically reviewed existing studies and identified gaps that needs further investigation. The findings highlight the need for more research on emerging social media platforms, the adoption of advanced AI-driven analytical models, the integration of social media intelligence into business decision-making, and the development of standardized evaluation frameworks. These gaps will guide the subsequent chapters in formulating a research approach that contributes to both academic knowledge and practical business applications.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research methodology, structured around the Data Science Trajectories (DST) model, which provides an iterative and adaptable framework for analyzing dynamic datasets. DST is particularly suitable for this research as it systematically guides the process of data acquisition, preparation, exploration, modeling, and evaluation, ensuring a structured yet flexible approach to analyzing TikTok influencer activity. The chapter begins by introducing the DST framework and its relevance to this research. It then outlines the business problem, focusing on the impact of TikTok influencers within the beauty and personal care industry. The subsequent sections detail the methodological workflow, covering data acquisition, preparation, and exploration, including the extraction of influencer revenue data from Kalodata and TikTok video/comment metadata from Apify. To provide a clear visual representation of the research workflow, Figure 3.1 maps the methodology, linking each phase to the corresponding research objectives and deliverables. This structured approach ensures alignment with the research goals, particularly in conducting descriptive analytics, sentiment analysis, and topic modeling. Finally, the chapter concludes by setting the stage for further modeling and evaluation, ensuring that the research objectives are met through rigorous data processing and analytical techniques.

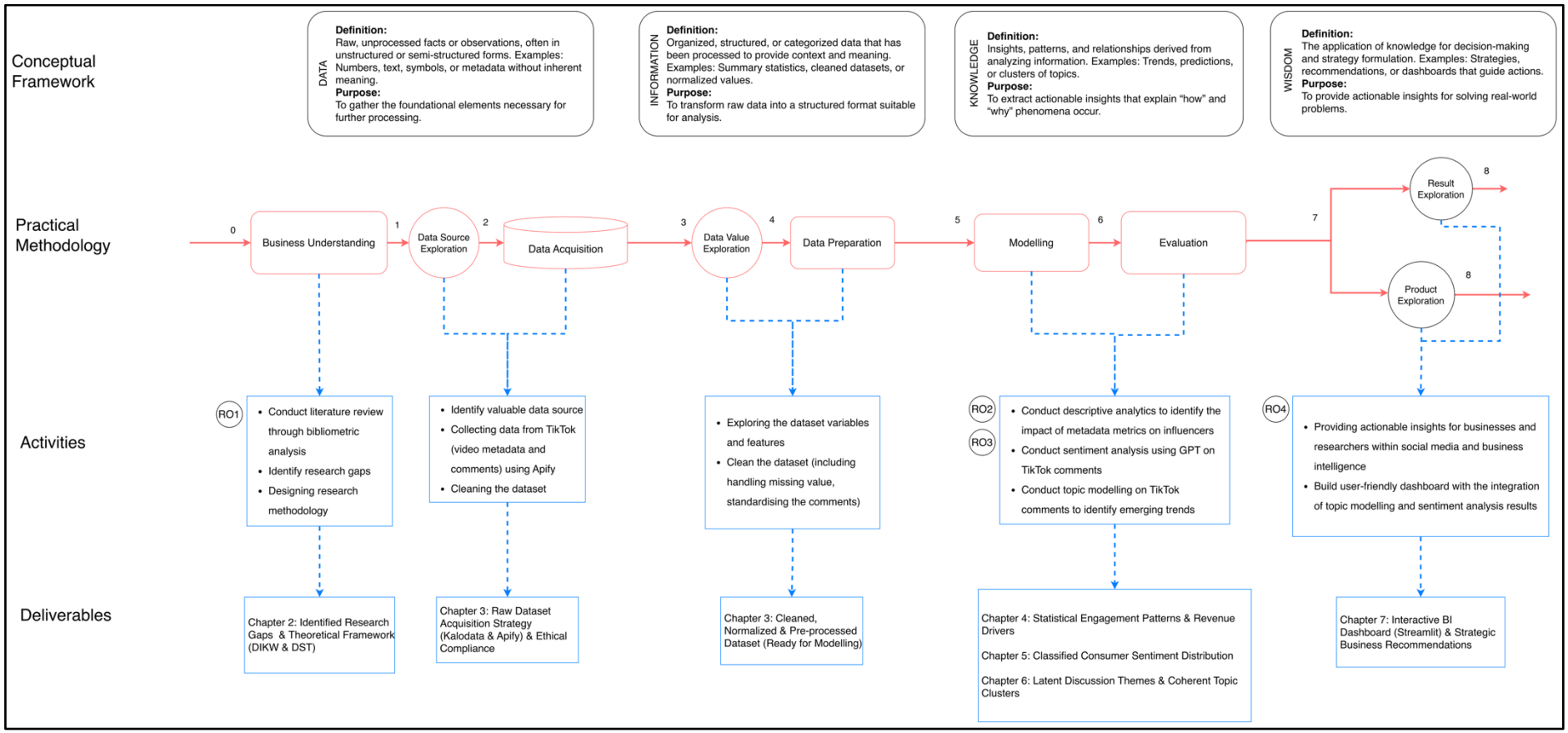


Figure 3.1 The Overview Of Research Methodology

3.2 Business Understanding

Understanding the business context is a crucial first step in this research, as it ensures that the methodological approach aligns with industry challenges and research objectives. In Chapter 2, a detailed literature review and bibliometric analysis established the landscape of TikTok influencer marketing, consumer engagement, and business intelligence. These insights highlighted key research gaps, particularly in understanding engagement patterns, sentiment trends, and content themes on TikTok. The findings from Chapter 2 directly influence the methodology presented in this chapter. Specifically, they shape data source selection, modeling approaches, and analytical techniques. For instance, the need to analyze engagement drivers led to the inclusion of descriptive analytics on metadata, while gaps in sentiment understanding justified the use of GPT-based sentiment analysis. This section provides a brief overview of the business context; however, a more detailed discussion is provided in Chapter 2. The next section, Data Source Exploration, outlines the datasets and tools used to operationalize these research insights

3.3 Data Source Exploration

Exploring various social media platforms helps identify which sources provide valuable insights into business-influencer interactions and customer behaviour. Established platforms like Facebook, YouTube, and Instagram have long offered rich datasets, shedding light on user engagement and strategy optimization for businesses (Li, 2022). This is reflected in the literature review, where terms like Twitter and Facebook frequently dominate discussions. However, newer and valuable data sources need to be explored to uncover the latest sources of user-generated content (UGC). Based on the literature review, it is evident that newer platforms like TikTok remain underexplored in the field of social media and business intelligence through the non-existence of the term "*TikTok*". As such, this research emphasizes TikTok, a rapidly growing platform with high user engagement rates, making it an ideal data source to examine emerging trends in digital marketing and consumer behaviour.

TikTok's appeal lies in its highly engaging, creative environment, where users interact with brands in ways not commonly seen on older platforms. The platform not only facilitates unique user engagement but also amplifies electronic word of mouth

(eWOM). On TikTok, users' comments, shares, and reactions play a significant role in shaping public perception of brands. This form of eWOM is particularly relevant in the beauty and personal care sector, where influencers and brands drive trends with visual and interactive content. TikTok's short-form video format and user-driven content ecosystem provide brands with an effective channel to reach audiences, gather real-time feedback, and foster community connections around products and trends.

Given TikTok's rising influence, particularly in beauty and personal care, this research uses TikTok as the primary data source. This choice allows for an in-depth exploration of UGC, including comments and videos, to assess how businesses and influencers in this sector leverage both UGC and eWOM to increase customer engagement, manage brand perception, and refine digital strategies.

3.4 Data Acquisition

Initially, the data understanding steps will start with collecting the required data to perform analysis. In this research, the data that will be used comes from two data sources which are Kalodata and Apify. These two platform provides a closely related data such as revenue information from Kalodata and TikTok metadata and comments from Apify. This data is crucial for this analysis as it forms the foundation for understanding influencer behaviour and their impact on brand engagement and revenue.

3.4.1 Kalodata

Kalodata is a powerful data analytics platform specializing in TikTok business analysis, with a focus on automatically calculating revenue generated by shops and influencers on TikTok. The platform offers multiple data views, enabling researchers to analyse revenue data based on products, creators, and categories. This flexibility allows for a comprehensive examination of various aspects of TikTok's commercial landscape, making it highly valuable for extracting business insights. Researchers and business professionals can leverage Kalodata's data analytics capabilities to streamline decision-making processes and enhance operational strategies, particularly in the domain of business intelligence.

Kalodata's effectiveness has been demonstrated in several academic studies. For example, Shabrina et al., 2024, utilized Kalodata to examine how the credibility of live-

streaming hosts influences consumer purchasing decisions, focusing on the Skintific brand. Their study highlights how the interaction between TikTok influencers and their audiences can drive business outcomes. Similarly, Yuli Wijayanti et al., 2024, investigated the influence of social media influencers' credibility, parasocial relationships, and brand image on consumer purchasing behaviour on TikTok, with the help of Kalodata's revenue tracking feature. These studies showcase Kalodata's role in providing critical revenue insights, offering a reliable dataset for both academic research and practical business applications. By integrating Kalodata into this research, a detailed understanding of TikTok's revenue-generating influencers is achieved, allowing for a deep analysis of influencer marketing dynamics within the beauty and personal care category.

Moreover, Kalodata's ability to provide detailed revenue insights across multiple perspectives, from individual products to broader business categories, further supports its usefulness in this research. A notable finding from Kalodata shows that the Beauty and Personal Care category recorded the highest revenue on TikTok in Malaysia, generating RM195.39 million in the 30 days leading up to September 28, 2024. This finding aligns with the research of Darmatama & Erdiansyah, 2021, which in a case study on beauty products, specifically lipstick, reported that 70.4% of respondents expressed interest in the product. The study emphasized that maintaining a positive brand image and ensuring high product quality are crucial factors that significantly enhance consumer interest and influence purchase decisions (Darmatama & Erdiansyah, 2021).

Additionally, influencer reviews and positive electronic word-of-mouth (eWOM) on TikTok have been shown to impact millennials' purchasing intentions, particularly in the beauty sector. For this research, the Beauty and Personal Care category was selected due to its potential to provide valuable insights into consumer preferences and trends, thereby offering opportunities for businesses to develop more effective marketing strategies. Within this category, the top 20 personal/individual accounts with the highest revenue were selected for further analysis. Figure 3.2 below illustrates the top 10 categories with the highest revenue.

	Category	Best-selling Products	Revenue	Revenue Growth Rate	Revenue(08/28 ~ 09/26)	Number of Shops	Revenue per Shop	Category Level
1	Beauty & Personal Care		RM195.39m	-25.85%		7237	RM27.00k	L1
2	Womenswear & Underwear		RM82.18m	-23.70%		5626	RM14.61k	L1
3	Muslim Fashion		RM72.48m	-6.77%		3466	RM20.91k	L1
4	Food & Beverages		RM69.91m	-19.50%		4087	RM17.10k	L1
5	Phones & Electronics		RM45.61m	-21.35%		3495	RM13.05k	L1
6	Sports & Outdoor		RM36.37m	-27.51%		4393	RM8.28k	L1
7	Home Supplies		RM32.79m	-16.49%		3596	RM9.12k	L1
8	Menswear & Underwear		RM30.39m	-21.59%		2857	RM10.64k	L1
9	Automotive & Motorcycle		RM29.01m	-27.80%		2446	RM11.86k	L1
10	Fashion Accessories		RM26.02m	-25.16%		3334	RM7.80k	L1

Figure 3.2 Top 10 Category With Highest Revenue

3.4.2 Apify

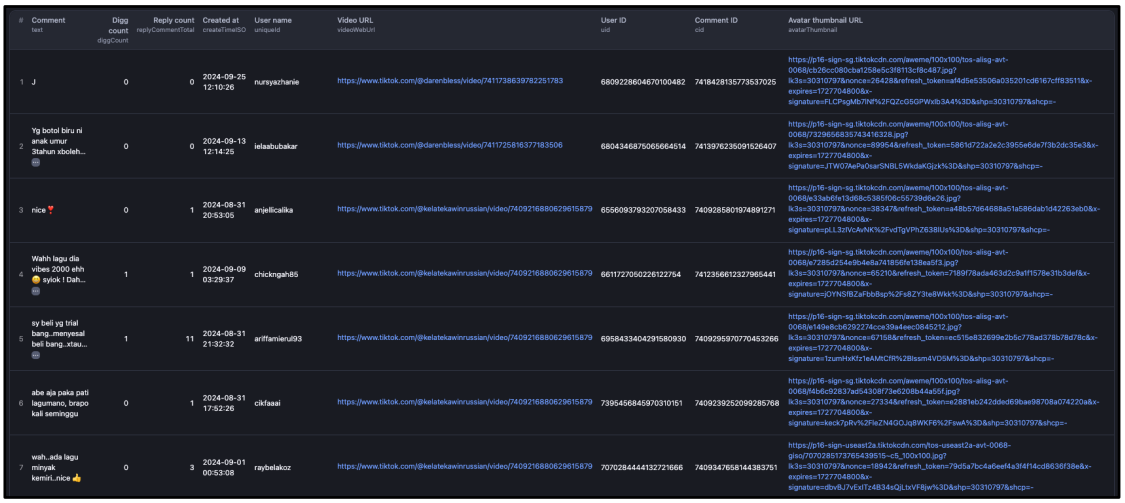
In this research, Apify serves as a vital tool for data extraction and web scraping, with a particular focus on TikTok data. Apify is renowned for its capability to extract structured data from a variety of social media platforms, including YouTube, Instagram, and TikTok, making it a powerful solution for large-scale data collection in academic and commercial research. Apify's tools allow for the extraction of video statistics (such as likes, comments, and shares), user profile data, and even trending topics or hashtags. This broad scope of functionality offers researchers a flexible and scalable approach to analysing social media platforms. One of the most prominent features of Apify is its Actor interface, which facilitates automated web scraping based on user-defined parameters, allowing for the systematic and efficient collection of data on a large scale. The platform has gained attention among researchers seeking to understand the dynamics of social media engagement, consumer behaviour, and content dissemination.

This platform has been widely used in academic studies. For example, B. B. Pereira & Ha, 2024 utilized Apify to gather over 3,840 TikTok videos, tagged with hashtags such as #climatechange and #sustainability, in a study examining misleading information related to environmental issues. Their research provided valuable insights into how misinformation spreads on TikTok and the key factors contributing to this

phenomenon. Another example, Tuğral et al., 2021 which leveraged Apify's scraping capabilities to collect 4,044 Instagram posts over a two-week period, which they then analysed to evaluate the accuracy and relevance of posts tagged with lymphedema-related hashtags. The study explored how social media users disseminate medical information, underscoring the role of platforms like Apify in understanding user-generated content. In the context of this research, Apify's automation tools allow for comprehensive and systematic data extraction, providing the framework for reliable and detailed insights into TikTok activity, particularly in the Beauty and Personal Care category. However, these platforms operate on a paid subscription basis, with pricing dependent on the plan selected. Apify offers various subscription tiers, with the free version providing a limited credit of $5 (RM21.07), and plans scaling up to $999 (RM4210.78) per month for greater access and data extraction capacity. The pricing models charge based on the volume of data collected, which is crucial for this research given the extensive amount of data needed for a robust analysis.

In this research, two key APIs from Apify were utilized for data collection. The first is the "*TikTok Profile Scraper*", which extracts detailed information on TikTok videos from specific profiles. Using data from Kalodata, the URLs of the top 20 TikTok influencer accounts in the Beauty and Personal Care category were identified. These URLs were then fed into the "*TikTok Profile Scraper*" API, which was configured to collect a maximum of 300 posts/videos per profile within a timeframe starting from January 1, 2024, to capture recent content. The actual number of videos collected varied across influencers, depending on their productivity in generating content. Ultimately, a total of 3,912 videos were extracted from these profiles, incurring a cost of $19.57 (RM82.49) based on the API's pricing structure of $5 (RM21.07) per 1,000 videos.

The data collection process was then continued to use the second API from Apify called "*TikTok Comments Scraper*". This API charges $1 (RM4.21) per 1,000 comments, and was set to gather up to 100 comments per posts/video. From the 3,912 videos initially collected, a total of 34,597 comments were retrieved, resulting in a cost of $34.60 (RM145.84) for the comment data. Analysing this huge volume of user-generated comments allows the research to capture consumer sentiment, preferences, and opinions regarding beauty and personal care products. Comments provide a rich dataset for understanding the effectiveness of influencer content in shaping purchasing decisions and driving engagement. Figure 3.3 below presents a snippet of the comments collected using Apify.



#	Comment text	Digg count diggCount	Reply count replyCommentTotal	Created at createTimeISO	User name uniqueId	Video URL videoWebUrl	User ID uid	Comment ID cid	Avatar thumbnail URL avatarThumbnail
1	J	0	0	2024-09-25 12:10:26	nursyazhanie	https://www.tiktok.com/@darenbless/video/7411738639782251783	6809228604670100482	7418428135773537025	https://p16-sign-sg.tiktokcdn.com/aweme/100x100/tos-alisg-avt-0068/cb26cc080cba1258e5c3f8113cf8c487.jpeg?lk3s=30310797&nonce=26428&refresh_token=aaf4d5e53506a035201cd6167cff83511&x-expires=1727704800&x-signature=FLCPsgMb7INf%2FQZcG5GPWxlb3A4%3D&shp=30310797&shcp=-
2	Yg botol biru ni anak umur 3tahun xboleh...	0	0	2024-09-13 12:14:25	ielaabubakar	https://www.tiktok.com/@darenbless/video/7411725816377183506	6804346875065664514	7413976235091526407	https://p16-sign-sg.tiktokcdn.com/aweme/100x100/tos-alisg-avt-0068/7329658833574316328.jpeg?lk3s=30310797&nonce=89954&refresh_token=5861d722a2e2c3955e6de7f3b2dc35e3&x-expires=1727704800&x-signature=JTW07kePa0sarSN8L5WkdaKGjzk%3D&shp=30310797&shcp=-
3	nice	0	1	2024-08-31 20:53:05	anjellicalika	https://www.tiktok.com/@kelatekawinrussian/video/7409216880629615879	6556093793207058433	7409285801974891271	https://p16-sign-sg.tiktokcdn.com/aweme/100x100/tos-alisg-avt-0068/e33ab6fa13d68c5385f06c55739d6e26.jpeg?lk3s=30310797&nonce=38347&refresh_token=a48b57d64688a51a586dab1d42263eb0&x-expires=1727704800&x-signature=pLL3zlVcAvNK%2FvdTgVPhZ63BlUs%3D&shp=30310797&shcp=-
4	Wahh lagu dia vibes 2000 ehh syiok ! Dah...	1	1	2024-09-09 03:29:37	chicknganh85	https://www.tiktok.com/@kelatekawinrussian/video/7409216880629615879	6611727050226122754	7412356612327965441	https://p16-sign-sg.tiktokcdn.com/aweme/100x100/tos-alisg-avt-0068/e7285d254e8bd4e8a741856fe128ea5f3.jpeg?lk3s=30310797&nonce=65210&refresh_token=7189f78ada463d2c9a1f1578e31b3def&x-expires=1727704800&x-signature=jOYNSfBZaFbb8sp%2Fs8ZY3te8Wkk%3D&shp=30310797&shcp=-
5	sy beli yg trial bang..menyesal beli bang..xtau...	1	11	2024-08-31 21:32:32	ariffamieruli93	https://www.tiktok.com/@kelatekawinrussian/video/7409216880629615879	6958433404291580930	7409295970770453266	https://p16-sign-sg.tiktokcdn.com/aweme/100x100/tos-alisg-avt-0068/e149e8cb6292274cce39a4eec0845212.jpeg?lk3s=30310797&nonce=67168&refresh_token=ec516e832699e2b5c778ad378b78d78c&x-expires=1727704800&x-signature=1zumHsKfzTeAMtCfR%2Blssm4VD5M%3D&shp=30310797&shcp=-
6	abe aja paka pati lagumano, brapo kali seminggu	0	1	2024-08-31 17:52:26	cikfaaai	https://www.tiktok.com/@kelatekawinrussian/video/7409216880629615879	7395456845970310151	7409239252099285768	https://p16-sign-sg.tiktokcdn.com/aweme/100x100/tos-alisg-avt-0068/f4b6c92837ad54308f73e6208b44a55f.jpeg?lk3s=30310797&nonce=27334&refresh_token=e2881eb242dded69baa98708a074220a8&x-expires=1727704800&x-signature=keck7pRv%2FleZN4GOJq8WKF6%2FswA%3D&shp=30310797&shcp=-
7	wah..ada lagu minyak kemiri..nice	0	3	2024-09-01 00:53:08	raybelakoz	https://www.tiktok.com/@kelatekawinrussian/video/7409216880629615879	7070284444132721666	7409347658144383751	https://p16-sign-useast2a.tiktokcdn.com/tos-useast2a-avt-0068-giso/7070285173765439515~c5_100x100.jpeg?lk3s=30310797&nonce=18942&refresh_token=79d5a7bc4a6eef4a3f4f14cd8636f38e&x-expires=1727704800&x-signature=dbvBJ7vExIT24B34sQjLtrVF8jw%3D&shp=30310797&shcp=-

Figure 3.3 The Snippets Of The Comments Collected Using Apify

3.4.3 Ethical Considerations and Data Compliance

Given the use of automated data collection tools (Apify), this research strictly adhered to ethical guidelines regarding user privacy and data usage. While the data collected from TikTok is publicly available, specific measures were implemented to ensure compliance with research ethics and to mitigate potential risks associated with user-generated content (UGC). The data collection process was confined strictly to publicly accessible data. The Apify actors employed (TikTok Profile Scraper and Comment Scraper) were configured to access only content that users had voluntarily published to the public domain. No private accounts, direct messages, or restricted content were accessed. This aligns with standard internet-mediated research practices where public social media posts are considered observable public behavior.

While automated data scraping exists in a complex legal landscape regarding Platform Terms of Service, this research operates under the principle of fair use for academic and non-commercial research. The data is used solely for statistical analysis and pattern recognition (aggregate sentiment and engagement trends) rather than for profiling individual users or commercial surveillance. Apify, as a platform, operates in compliance with GDPR and CCPA regulations by processing only public data, ensuring that the collection method does not bypass security protocols or access unauthorized information. To address ethical concerns regarding user privacy, a strict data anonymization protocol was implemented immediately post-collection.

i. Personally Identifiable Information: Direct identifiers such as user IDs and specific usernames were removed or hashed during the pre-processing stage before any sentiment analysis or topic modeling was conducted.

ii. Aggregate Reporting: Findings in this thesis are reported at an aggregate level (e.g., "70% of comments were positive"). Where individual comments are cited as examples, they are paraphrased or presented without attribution to specific users to prevent re-identification, ensuring the "dignity and autonomy" of the participants is preserved.

This research acknowledges the need for ethical clearance when dealing with human participants. However, as this study involves the observation of public behavior online without direct interaction with participants (non-interventional) and does not collect sensitive personal data (health records, financial data), it falls under the category of low-risk research. This research was also formally reviewed and granted ethical approval by the Research Ethics Committee of the Research Management Centre, Universiti Teknologi MARA.

3.5 Data Value Exploration

In this section, an overview of the variables used in this research is provided, each playing a significant role in understanding influencer performance and audience engagement on TikTok. The variables encompass various types of data such as text-based, numerical, and list-based. These variables cover both qualitative and quantitative aspect of the data which could provide a better understanding and deeper insights of influencer's popularity and business impact. Text-based variables such as caption_text and comment_text may offer a big advantage which can be analysed for sentiment and discovers the main themes among discussed among users. Additionally, the numerical variables such as followers, views, revenue, and productCount may offers the understanding on the reach and engagement levels of the influencer's videos. Engagement metric like likesCount, commentCount, shareCount, and savesCount are critical to assess how audiences interact with the content, leading to user preferences and content quality. Hashtags and hashtagCount will allow a better understanding towards the discoverability strategies used by influencers. A detailed description for each of the variables are shown in the table 3.3 below.

Table 3.1
Data Descriptions For Every Variables

Parameters	Descriptions	Data Type
Nickname	It is a display name chosen by the TikTok user	String
Followers	The number of users following a specific influencer	Integer
Views	The total number of views from every video for a specific influencer	Integer
Revenue	The total sales accumulating from live event and videos	Integer
ProductCount	The number of products promoted or sell by the influencers	Integer
LiveNum	The total number of live streams the influencer hosted	Integer
LiveGmv	The total sales generated from live events	Float
VideoNum	The total number of videos posted by the influencers	Integer
VideoGmv	The total sales generated from video's posted by the influencers	Float
id	A unique identifier for each TikTok videos	Integer
name	The name associated with each TikTok profiles	String
likesCount	The total number of likes received for a specific video	Integer
playCount	The total number of playbacks for each video	Integer
shareCount	The total number of times the video is being shared by TikTok users	Integer
commentCount	The total number of comments on a specific video	Integer
savesCount	The total number of times a video is saved by users for later viewing	Integer
hashtagCount	The number of hashtags used in a video or post	Integer
duration	The length of the video in seconds	Float
hashtags	The list of hashtags name associated with a video or post	List (String)
caption_text	The text description that influencer writes with video or post.	String
comment_text	The text of comments made by users on TikTok videos	String

Basic data from TikTok, regarding the category of 'Beauty & Personal Care' influencers, is bound to be very helpful for business. It will enable brands to understand how various influencer channels are doing by taking a sneak peek look at key factors such as followers, views, and revenue. A good example is that of comparison, it can highlight the influences that are strongly related to their followers. This can be done by comparing, for instance, the number of followers with that of views. This helps in the identification of those influencers who, besides attracting people, also retain them. The analysis of lucrative data from such campaigns helps businesses find which of the influencers is best at converting views to actual sales, enabling them to concentrate their marketing efforts on the most profitable partnerships.

Also, data might reveal some kind of trends related to the kind of products in demand and those influencers who have probable potential for marketing specific kinds of products well. It keeps them ahead of the industry, making sure to get wind of an emerging trend well before their competitors even know such a thing exists. This helps in smarter campaign planning whereby businesses can pick only those influencers whom they know have previously been able to promote similar products with proven success. Also, segmenting the influencers concerning their performance-hence, the small but more engaging ones versus the larger ones that would target broader appeal-helps a company in targeting correctly. This will also help in measuring the effectiveness of marketing campaigns and useful for making strategies based on real performances. By looking at what return on investment different influencers provide, a business will know where to invest its marketing dollar. Because of that, from such insights into data, a brand can monitor competitive positioning and find new avenues in collaborating or partnering with influencers that best suit their brand image. With that, data in TikTok could be one of the powerful grounds for better marketing strategies down the line with ensured sales deal in business related to beauty and personal care.

3.6 Data Preparation

In this section, the data gathered from Kalodata and Apify will undergo a comprehensive data cleaning process. This involves selecting key variables, adjusting data types, and using a Generative Pre-trained Transformers (GPT) model to classify TikTok account types, ensuring proper categorization between personal and official brand accounts. Additionally, the GPT-4o model will be applied to clean and correct

comment text, addressing issues like slang and abbreviations to enhance data consistency. Finally, the datasets from both platforms will be merged to provide a comprehensive view of how influencer activities affect product sales in the Beauty and Personal Care sector, laying the foundation for subsequent analysis.

Initially, there are three primary datasets which are revenue data from Kalodata, videos from TikTok accounts (via Apify), and comments from the selected TikTok videos (via Apify). These datasets contain variables that vary in relevance and structure, necessitating the extraction of specific data and transformation of certain fields to ensure compatibility across all datasets. This stage of data preparation begins with selecting only the relevant variables required to address the research questions. For example, from the revenue dataset, variables like “Rank”, “Nickname”, “Followers”, “Revenue”, “ProductCount”, “LiveNum”, “LiveGmv”, “VideoNum”, “VideoGmv”, and “Views” were identified as key variables that provide insights into influencer activity and its effect on sales. These selected variables focus on capturing the key metrics associated with influencer performance and engagement. Below is the python code for reading and selecting the relevant columns from the dataset.

```
revenue_dataset = pd.read_csv('/content/drive/MyDrive/Akmal-
ResearchMaster/Advanced_TikTok_Analysis/updated_csv_files/updated_tiktok_dataframe.csv')
revenue_dataset = revenue_dataset [['Nickname', 'Followers', 'Revenue', 'ProductCount', 'LiveNum',
'LiveGmv', 'VideoNum', 'VideoGmv', 'Views', 'TiktokUrl']]
revenue_dataset.head()
```

Python code snippet to select variables from revenue dataset

Next, in the data preparation pipeline, the use of the GPT model was essential for classifying TikTok account types. Each account in the dataset needed to be classified as either “Individual/Personal” or “Organizational/Company” to facilitate a more detailed analysis of influencers. The GPT model was employed to automate this classification based on profile characteristics, ensuring the proper categorization of accounts. A Python function was developed to handle the API calls to OpenAI, feeding TikTok profile data into the GPT model for classification. The Python code and the sample output of this process are detailed below and Table 3.2, respectively.

```
def classify_tiktok_profile(profile_url):
   prompt = f"""
            Your task is to analyze the following tiktok profile: {profile_url},
            and decide whether it is an individual/personal account or
            organization/company account.
            Your response MUST ONLY 'Individual/Personal' or 'Organization/Company'."""

   try:
     response = client.chat.completions.create(
            model="gpt-4o",
            messages=[
               {"role": "system", "content": "You are professional and expert in TikTok Analytics"},
                    {"role": "user", "content": prompt}
               ],
           )

     classification = response.choices[0].message.content.strip()

     return classification

   except Exception as e:
       print(f"Error analyzing profile {profile_url}: {e}")

       return None

revenue_dataset.loc[:, "Profile_Type"] = revenue_dataset ["TiktokUrl"].apply(classify_tiktok_profile)
```

Python code utilising GPT to classify account type

Table 3.2
The Result Sample After Classifying The Tiktok Accounts

Followers	Revenue	Product Count	No. of Live	Live GMV	No. of Video	Video GMV	Views	Profile Type
55,320	116,410	10	155	86,250	37	30,150	1,130,000	Organization/ Company
310,300	97,730	5	267	96,790	35	937	754,160	Individual/ Personal
147,660	97,320	40	79	74,340	6	22,980	300,870	Organization/ Company
14,500	84,940	4	66	82,960	10	1,980	329,620	Organization/ Company
181,930	73,450	13	18	52,740	6	20,720	1,750,000	Individual/ Personal

Additionally, the revenue dataset underwent further transformation to create new variables that could provide more insightful metrics. These included "Revenue per Follower," "Revenue per Video," and "Revenue per Live," which were calculated by dividing the total revenue generated by the respective number of followers, videos, and

live sessions. These new variables allow for a deeper understanding of the type of content or activity that generates the most revenue for influencers, which is crucial for identifying trends and patterns in influencer marketing within the beauty industry. The code for generating these new variables is illustrated below, and the results are shown in Table 3.3.

```
revenue_df[“RevenuePerFollower”] = (revenue_df[“Revenue”] / revenue_df[“Followers”]
revenue_df[“RevenuePerVideo”] = revenue_df[“VideoGmv”] / revenue_df[“VideoNum”]
revenue_df[“RevenuePerLive”] = revenue_df[“LiveGmv”] / revenue_df[“LiveNum”]
```

Creating new variables based one revenue data

Table 3.3
The Sample Results Of Creating The New Variables

Followers	Revenue	No. of Live	Live GMV	No. of Video	Video GMV	Revenue Per Follower	Revenue per Video	Revenue Per Live
394,680	246,150	27	246,150	18	0	0.62	0.00	9,116.67
63,480	293,010	61	110,020	42	182,990	4.62	4,356.90	1,803.61
54,150	311,150	235	130,750	25	180,390	5.75	7,215.60	556.38
63,800	348,890	56	317,370	47	31,520	5.47	4,502.86	5,667.32
132,950	73, 191,460	7	0	44	191,460	1.44	4,351.36	0.00

The next step in the data cleaning process was the treatment of the video and comments datasets extracted from Apify. The video dataset contained numerous variables, but only variables such as “*id*”, “*authorMeta/name*”, “*authorMeta/nickname*”, “*createTimeISO*”, “*diggCount*”, “*playCount*”, “*shareCount*”, “*commentCount*”, “*collectCount*”, are selected for analysis. The “*createTimeISO*” variable, which recorded the timestamp of each video, was reformatted to include only the day, month, and year (DD-MM-YYYY). The “*diggCount*” variable, representing the number of likes, was renamed to “likesCount” for clarity. Similarly, hashtags were also processed because Apify’s data structure separates hashtags into individual variables like “*hashtag/0*”, “*hashtag/1*”, and it goes until “*hashtag/21/name*”, so the total number of hashtags per video was counted and stored in a new variable called “*hashtagCount*”. All hashtags for each video were then combined into a single variable, with each hashtag separated by a semicolon. These transformations ensured that the data was both clean and structured for effective analysis. The Python code for these

steps is outlined below, and Table 3.4 presents sample results of the cleaned video dataset.

```
ori_df = pd.read_csv('/content/drive/MyDrive/Akmal-ResearchMaster/TikTok
Analysis/latest_dataset/dataset_tiktok-profile-scraper_2024-09-29_10-44-10-350.csv')

# Convert the createTimeISO format
ori_df['createTimeISO'] = pd.to_datetime(ori_df['createTimeISO'], format="%Y-%m-%dT%H:%M:%S.%fZ")
ori_df['createTimeISO'] = ori_df['createTimeISO'].dt.strftime("%d-%m-%Y")

# Selecting columns to perform analysis
columns = ['id','authorMeta/name','authorMeta/nickName','diggCount', 'playCount', 'shareCount',
'commentCount', 'collectCount','videoMeta/duration', 'text', 'hashtags/0/name', 'hashtags/1/name',
'hashtags/2/name', 'hashtags/3/name','hashtags/4/name', 'hashtags/5/name', 'hashtags/6/name',
'hashtags/7/name', 'hashtags/8/name', 'hashtags/9/name', 'hashtags/10/name', 'hashtags/11/name',
'hashtags/12/name', 'hashtags/13/name','hashtags/14/name', 'hashtags/15/name', 'hashtags/16/name',
'hashtags/17/name', 'hashtags/18/name','hashtags/19/name', 'hashtags/20/name', 'hashtags/21/name']

# Include the selected column into main df
main_df = ori_df[columns]

# Rename the column
main_df.rename(columns={'authorMeta/name': 'name', 'authorMeta/nickName': 'nickname', 'diggCount':
'likesCount', 'collectCount': 'savesCount', 'text': 'caption_text', 'videoMeta/duration':
'duration'}, inplace=True)

# Selecting all hastag column
hashtag_col = ['hashtags/0/name', 'hashtags/1/name', 'hashtags/2/name', 'hashtags/3/name',
'hashtags/4/name', 'hashtags/5/name', 'hashtags/6/name', 'hashtags/7/name',
'hashtags/8/name','hashtags/9/name', 'hashtags/10/name', 'hashtags/11/name', 'hashtags/12/name',
'hashtags/13/name','hashtags/14/name', 'hashtags/15/name', 'hashtags/16/name', 'hashtags/17/name',
'hashtags/18/name','hashtags/19/name', 'hashtags/20/name', 'hashtags/21/name']

# Counting the number of hashtag
video_df['hashtagCount'] = video_df[hashtag_col].apply(lambda row: row.notna().sum(), axis=1)

# Combining all hashtags into a single column
video_df['hashtags'] = video_df[hashtag_col].apply(lambda row: ';'.join(row.dropna().astype(str)),
axis=1)

# dropping the old hashtag columns
video_df.drop(columns=hashtag_col, inplace=True)
```

Python code of extracting and renaming the selected variables

Table 3.4
Sample Results Of The Selected Variables

ID	...463424	...484416	...889984
Date Created	29-04-2024	03-03-2024	06-09-2024
Likes Count	671	10,600	602
Play Count	45,200	165,500	28,600
Share Count	30	78	55
Comment Count	24	58	139
Save Count	176	608	79
Duration	33s	109s	31s
Caption Text	Video testing di sini...	Raya nanti makeup...	Replying to @...
Hashtag Count	1	1	0
Hashtags	personal	beauty	None

Engagement rate is an important measure in social media marketing that shows how well content performs on social media platforms like TikTok, Instagram, and Facebook. It helps influencers and marketers see how much their audience interacts with their posts through likes, comments, shares and save (Monacho & Slamet, 2023). When a post receives a high level of interaction, it increases brand visibility. This metric also offers insights into audience preferences by highlighting which posts get the most or least interaction. For this research, the engagement rate was calculated by summing the number of likes, shares, and comments and dividing this total by the number of video views. This calculation in python code is illustrated below, while Table 3.5 provides a sample output showing the engagement rate for a subset of videos.

```
# Calculating the engagement rate
final_df['engagement_rate'] = (final_df['likesCount'] + final_df['commentCount'] +
final_df['shareCount'] + final_df['savesCount']) / final_df['playCount']
```

Python code of calculating the engagement rate

Table 3.5
The Sample Result Of The Engagement Rate Calculation

ID	Likes Count	Comment Count	Share Count	Saves Count	Play Count	Engagement Rate
...29503744	178	23	4	15	8,178	0.027146
...69934336	1,697	25	33	63	76,500	0.023739
..49202176	278,300	440	1,136	9,975	4,200,000	0.069012

The comment dataset, also extracted using Apify, required its own set of cleaning procedures. Variables such as "text" (which contained the comment text) and "*diggCount*" (which counted the number of likes for each comment) were renamed to "*comment_text*" and "*commentLikesCount*" to clarify their meaning. Additionally, the video URL associated with each comment needed to be processed to extract the video ID, which was then stored in a new variable called "*video_id*". Since each video could have multiple comments, the data was grouped by *video_id*, with the comments for each video being combined into a single entry, separated by semicolons. Meanwhile, the likes for each comment were summed to create a total "*commentLikesCount*" for each video. This cleaned comment dataset was then merged with the video dataset using the "*video_id*" as the unique identifier, resulting in a single dataset that included both video metadata and the corresponding user-generated comments. The steps for this process are outlined below, with sample results presented in Table 3.6.

```
# Function to extract video id from videoWebUrl
def extract_video_id(text):
    match = re.search(r'/video/(\d+)', text)

    if match:
      # Return the first matched group (the video ID)
      return match.group(1)

    # Return None if no match is found
    return None

# Renaming the the text variable
comment_df = comment_df.rename(columns={'text': 'comment_text', 'diggCount': 'commentLikesCount'})

# Extracting the video id from videoWebUrl
comment_df['video_id'] = comment_df['submittedVideoUrl'].apply(extract_video_id)
comment_df['video_id'] = comment_df['video_id'].apply(lambda x: int(float(x)))
comment_df['video_id'] = comment_df['video_id'].astype('int')

# Combining multiple comments with the same video_id
comment_df = comment_df.groupby('video_id').agg({
      'comment_text': lambda x: ';'.join(x.astype(str)),  # Group the comment_text with semicolon
      'commentLikesCount': 'sum'  # Sum commentLikesCount
  }).reset_index()

# Merge the comment_text and commentLikesCount into main df
final_df = pd.merge(video_df, comment_df, left_on='id', right_on='video_id', how='inner')
final_df.drop(columns=['video_id'], inplace=True
```

Python code of processing the comment dataset

Table 3.6
Sample Results Of Processing The Comment Dataset

ID	Date Created	Caption Text	Comment Text
…88480	22-08-2024	Akak tengah packing order ada orang puji pulak🥰	Maniss nya akak… 🤣 MANIS BANYAKKKKK…
…30528	12-09-2024	feedback terlampau banyak korang kakSue je tak...	Dari Singapura bleh order?...
…16000	17-07-2024	step son where u at 🤣	So pretty; Love only meeee…

The data cleaning pipeline continues with cleaning the TikTok comment text by leveraging the GPT-4o model to standardize the vocabulary and reduce the overall dimensionality of the text (Hickman et al., 2022). This step is crucial as it helps address the amount of misspelled words, slang, and short-form abbreviations that are commonly found in user-generated content on social media in order to achieve high-quality text analysis (Batrinca & Treleaven, 2015). These variations in language can significantly impact the analysis, which could lead to potential misinterpretations of sentiment, topics, and meaning. The GPT-4o model works with prompts, so in this research, a prompt is carefully crafted to correct every comment text. The cleaning instruction that is given to GPT-4o is to make sure every word is in Malay and correct any misspelled words, slang, and short-form abbreviations. It is intentionally that special characters and emoji are not removed from the text, as they might carry the tone or support sentiment meanings. Below shows the code to utilize GPT to clean the text along with its prompt, and Table 3.7 shows the sample result of cleaned comment text, highlighting the transformation of the comment from its raw, user-generated form to a standardized version that is ready for further analysis.

```
# Function to clean the comment text
def clean_comment(text):
    prompt = f"""
    There are English words in the following text. Translate the words to MALAY.
    Return ONLY THE TRANSLATED COMPLETE TEXT. ALSO FIX SOME SHORTFORM/SLANG/MISSPLELLED WORD INTO
NORMAL MALAY WORDING: {text}
    """

    response = client.chat.completions.create(
                model="gpt-4o",
                messages=[
                    {"role": "system", "content": "You are an expert multilingual text analysis."},
                    {"role": "user", "content": prompt}
                ],
            )

    return response.choices[0].message.content.strip()

# Apply the cleaning function to the 'comment_text' column and store the result in a new column
final_df['cleaned_comment_text'] = final_df['comment_text'].apply(clean_comment)
```

Python code to instruct GPT to clean comment text

Table 3.7
The Sample Result Of Before And After GPT-4o Clean The Comment Text

Comment Text	Cleaned Comment Text
happy to see this niena🥰🥰🥰 and happy for you. ...	gembira melihat ini niena🥰🥰🥰 dan gembira untuk...
sy setiap malam tiap kali tdokn baby pasti aka...	saya setiap malam tiap kali tidurkan bayi past...
result please🥺;Wow;Result!;Best;Partt 2 pls 😭;...	keputusan tolong 🥺;Wow;Keputusan!;Terbaik;Baha...

This comprehensive data cleaning process not only enhanced the quality of the dataset but also ensured that it was ready for further analysis. By addressing inconsistencies, removing irrelevant variables, and standardizing text, this research lays the groundwork for accurate and insightful analysis of how influencer activities on TikTok impact consumer behaviour and sales in the Beauty and Personal Care sector.

For the subsequent phases which is the modeling phase is broken down into specific chapters to allow for detailed exploration of each analytical approach. Chapter 4 focuses on Descriptive Analytics, where key statistics are analysed to understand overall trends and patterns in the data. Chapter 5 delves into Sentiment Analysis, applying both general and specialized algorithms to gauge consumer sentiment in TikTok comments. Chapter 6 covers Topic Modeling, using advanced techniques to

uncover key themes from the data. The evaluation phase for each model is conducted within its respective chapter to assess the performance and insights drawn from each analysis. Further the methodology phase, the result and product exploration phase are discussed in chapter 7.

3.7 Conclusion

In conclusion, Chapter 3 has outlined the comprehensive methodology for this research, guided by the DST model. The process started with a thorough understanding of the business problem and exploration of key data sources, namely Kalodata and Apify. The data preparation phase ensured the datasets were ready for detailed analysis, which will be carried out in the subsequent chapters. Each dataset will be examined through specific modeling techniques, including descriptive analytics, sentiment analysis, and topic modeling, with evaluations conducted within their respective chapters. This structured approach will help uncover valuable insights into TikTok's influence in the beauty and personal care industry.

CHAPTER 4
DESCRIPTIVE ANALYTICS

4.1 Introduction

In this research, descriptive analytics for both datasets prepared in the previous chapter is separated into two subsections which are Revenue Analytics and Metadata Analytics. Both analytics will first undergo basic analytics such as summary statistics, and further the analytics into more specific analytics based on each dataset's variables. This analysis performed is to provides a foundational understanding on the revenue's and TikTok video's data.

4.2 Exploratory Data Analysis (EDA)

Table 4.1 illustrates the exploratory data analysis of the TikTok dataset. The report provides a fascinating snapshot of the performance and scale of 20 TikTok accounts over a 30-day period (August 28 to September 26). These accounts collectively generate significant revenue, totalling approximately RM9.3 million, which breaks down to an average of about RM465,000 per account. This high average suggests that each account is achieving notable financial success, potentially driven by a strong follower base and active engagement. The accounts collectively offer 681 different products, averaging 34 products per account, indicating a diverse range of offerings that might cater to varied consumer interests.

The collective followers of these accounts is substantial, with a total of 18,383,760 followers, which equals to an average of 919,188 followers per account. This high follower counts per account points to a strong influence and reach within the Beauty & Personal Care industry. Furthermore, the total number of views across all accounts is 182,346,160 views, with each account accumulating an average of 9,117,308 views. It's clear that these TikTok beauty accounts hold significant popularity and influence. The engagement through likes is also noteworthy, with a total of 63,051,709 likes and an average of 1,611,710 likes per account, illustrating strong level of engagement and interaction from the audiences.

Moreover, the play count is extremely high with a total of 1,326,714,387 plays, giving an average of 339,139 plays per account, reflecting consistent viewer interest and repeated content consumption. Shares, which usually indicate content virality, amount up to 1,645,977 in total, with an average of 82,299 shares per account, underscoring the strong word-of-mouth or electronic word-of-mouth (eWOM) impact. Sharing is a crucial aspect since it reflects users' willingness to recommend the content to others, thus amplifying the reach beyond the existing follower base. Overall, this data reinforces the effectiveness of TikTok as a platform for Beauty and Personal Care brands to engage audiences and drive sales. High numbers in followers, views, and engagement metrics (likes and shares) reflect the potential for brands to harness this platform's massive reach and engagement capacity to enhance visibility, foster community, and boost revenue.

Table 4.1
Exploratory Data Analysis Report

Parameters	Details
Number of TikTok account	20
Total 30-Day Revenue (Aug 28 – Sep 26) (MYR)	9,298,430
Average Revenue (MYR) per Account	464,921.50
Total number of Products	681
Average number of Products per Account	~34
Total Number of Followers	18,383,760
Average number of Followers per Account	919,188
Total Number of Views	182,346,160
Average Number of Views per Account	9,117,308
Total Number of Likes	63,051,709
Average Number of Likes per Account	16,117.51
Total Number of Play Count	1,326,714,387
Average Number of Play Count per Account	339,139.70
Total Number of Share Count	1,645,977
Average Number of Share Count per Account	420.75

4.3 Revenue Analytics

Revenue analytics kick off with a summary statistic on the revenue-related variables. The summary statistic encompasses the calculation of the mean, median,

standard deviation, minimum, and maximum for metrics such as “Followers”, “Views”, “Revenue”, “Product Count”, “Live Num”, “LiveGmv”, “Video Num”, “VideoGmv”, “Revenue per Follower”, “Revenue per Video”, and “Revenue per Live”. These statistics could offer insights into the central tendencies and variability of the revenue data. To achieve the calculation for the metrics, a python library named SciPy is employed. The python codes to calculate the statistics is illustrated below while the summary statistics results are shown in the table 4.2.

```
# List of the metrics
metrics = ['Followers', 'Views', 'Revenue', 'ProductCount', 'LiveNum', 'LiveGmv', 'VideoNum',
'VideoGmv','RevenuePerFollower', 'RevenuePerVideo', 'RevenuePerLive']

# Create an empty dictionary to store the results
statistics = {}

for metric in metrics:
    mean_val = round(revenue_analytics_df[metric].mean(), 2)
    median_val = revenue_analytics_df[metric].median()
    mode_val = revenue_analytics_df[metric].mode()[0]  # Taking the first mode in case of multiple
modes
    min_val = revenue_analytics_df[metric].min()
    max_val = revenue_analytics_df[metric].max()
    std_val = round(revenue_analytics_df[metric].std(), 2)

    # Store the results in the dictionary
    statistics[metric] = {
        'Mean': mean_val,
        'Median': median_val,
        'Mode': mode_val,
        'Min': min_val,
        'Max': max_val,
        'Standard Deviation': std_val
    }

# Convert dictionary to DataFrame for better visualization
statistics_df = pd.DataFrame(statistics)

# Format the statistics to 2 decimal places
statistics_df[metrics] = statistics_df[metrics].applymap('{:.2f}'.format)

# Transpose the DataFrame to have metrics as rows and stats as columns
statistics_df = statistics_df.T

# Print the results
statistics_df
```

Python code to calculate the summary statistics

Table 4.2
Summary Statistic Of Revenue Data

Parameter	Mean	Median	Mode	Min	Max	Standard Deviation
Followers	591904.3	167120.0	9210.0	9210.0	5700000.0	1157825.0
Views	6995796.6	4270000.0	1660000.0	177800.0	28390000.0	7885817.9
Revenue	307492.2	250600.0	172480.0	172480.0	1540000.0	204883.6
Product Count	24.4	11.0	5.0	1.0	174.0	35.7
LiveNum	60.1	39.0	0.0	0	256.0	67.4
LiveGmv	144806.1	125945.0	0.0	0.0	510860.0	136146.6
VideoNum	36.8	22.0	1.0	0.0	187.0	40.8
VideoGmv	162786.4	172420.0	0.0	0.0	1180000.0	184475.4
Revenue Per Follower	2.9	1.6	0.2	0.04	20.4	3.5
Revenue Per Video	19740.6	4698.2	0.0	0.0	236000.0	40312.05
Revenue Per Live	3182.5	1380.6	0.0	0.0	17403.3	4067.3

The number of followers and views are key indicators of an influencer's reach on the platform. Based on the summary statistic, the average influencer has about 591,904 followers, but the median is much lower at 167,120. This suggests that while a few influencers have millions of followers, most have fewer. The smallest follower count is 9,210, and the largest is 5,700,000. Similarly, views range widely from 177,800 to 28,390,000, with an average of about 6,995,797 and a median of 4,270,000. These figures suggest that while there is potential for massive reach on TikTok, achieving such levels is relatively rare, and most influencers have a more modest audience size. To attract more followers, influencers with smaller audiences should focus on creating engaging and high-quality content. Trying out different styles and formats can help determine what appeals most to their target viewers. Engaging directly with the audience by responding to comments and messages can build a loyal community, leading to organic growth in followers and views.

Revenue reflects how well influencers turn their reach into earnings. The average revenue among these influencers is RM307,492, but the median is RM250,600. Earnings range from RM172,480 to RM1,540,000, indicates that there is a big difference between the highest and lowest earners. This inequality suggests that while

some influencers earn substantially more than others, the majority earn closer to the median value. The wide range and high variation point to considerable differences in earning potential among influencers. To increase revenue, influencers should consider diversifying their marketing strategies. For instance, setting realistic revenue goals based on the median earnings can help influencers track their progress. Studying the strategies of top earners might reveal effective practices that others can adopt.

Creating content is essential for attracting and keeping an audience. On average, influencers promote about 24 products, with some promoting as few as one and others as many as 174. They host an average of 60 live sessions and post about 37 videos, though some do none at all, and others produce a lot more. These differences highlight that influencers employ diverse content strategies. If an influencer is posting fewer videos or live sessions than average, increasing the frequency could keep their audience more engaged. However, it is important to maintain high quality to ensure the content remains valuable. Live sessions can be especially effective for building a stronger connection with followers. When promoting products, focusing on items that closely related to audience interests can enhance the effectiveness.

Earning money from live sessions and videos can significantly impact an influencer's total revenue. The average revenue from live sessions is RM144,806, with some influencers earning nothing and others earning up to RM510,860. For videos, the average revenue is RM162,786, ranging from RM0 to RM1,180,000. These figures reveal that while there is significant earning potential from both live sessions and videos, not all influencers take advantage of these opportunities. Influencers not earning from live sessions should explore ways to monetize them, such as product placements or sponsored content. Engaging viewers during live sessions with interactive features can encourage participation and increase earnings. For posted videos, analysing which ones generate the most income can help influencers understand what works best. They can then focus on creating similar content to maximize revenue. Experimenting with different types of content such as tutorials, reviews, or behind-the-scenes footage may also help identify what drives higher earnings.

Efficiency metrics show how well influencers convert their audience and content into revenue. The average revenue per follower is RM2.98, but this ranges from as low as RM0.04 to as high as RM20.43. For revenue per video, the average is RM19,741, with a wide range from RM0 to RM236,000. These efficiency metrics highlight that not all influencers are equally successful in monetizing their content and audience. If an

influencer's revenue per follower is low, they might introduce exclusive content, offer promotions, or engage in affiliate marketing to boost earnings from their existing audience. Improving the value of content by making it more engaging can encourage viewers to take actions that increase revenue, like making purchases or sharing the content. Regularly reviewing these efficiency metrics can help influencers spot trends and adjust their strategies accordingly. If certain types of content or approaches yield better returns, focusing more on those areas can enhance overall performance.

Next, the analytics continues to examine the correlation matrix to understand the relationship between different revenue-related variables. Pandas, a python library which equipped with a convenient library to calculate the Pearson correlation coefficients, by this, the strength and direction of the relationship between variables such as “Followers”, “Views”, “Revenue”, “Product Count”, “Live Num”, LiveGmv”, “Video Num”, VideoGmv”, “Revenue per Follower”, “Revenue per Video”, and “Revenue per Live” can be identified. This step is crucial for isolating factors most strongly associated with revenue, thereby providing valuable insights for strategic decision-making. To ensure an objective and consistent interpretation of these statistical results, the coefficients are categorized based on the standardized scale of strength presented in Table 4.3 which adapted from (Schober et al., 2018). It is important to note that while these ranges provide a conventional framework, statistical literature suggests they should be used as guidelines rather than strict rules, as the practical significance of a correlation often depends on the specific scientific or business context (Schober et al., 2018; Taylor, 1990). Below is the Python code used to compute these coefficients.

```
# Calculate the correlation between duration and other metrics
correlations = revenue_analytics_df[['Followers', 'Views', 'Revenue', 'LiveNum','LiveGmv',
'VideoNum', 'VideoGmv']].corr()

# Display correlation coefficients
correlations
```

The python code to calculate the correlation coefficients

Table 4.3
Interpretation Of Pearson Correlation Coefficients (*r*)

Correlation range (*r*)	Interpretation	Description
0.70 to 1.00	Strong/High	A very predictable relationship where variables move closely together.
0.40 to 0.69	Moderate	A clear and consistent trend is visible between the variables.
0.10 to 0.39	Weak/Low	A slight relationship exists, but other factors are likely more influential.
0.00 to 0.09	Negligible	Little to no linear relationship is detected between the variables

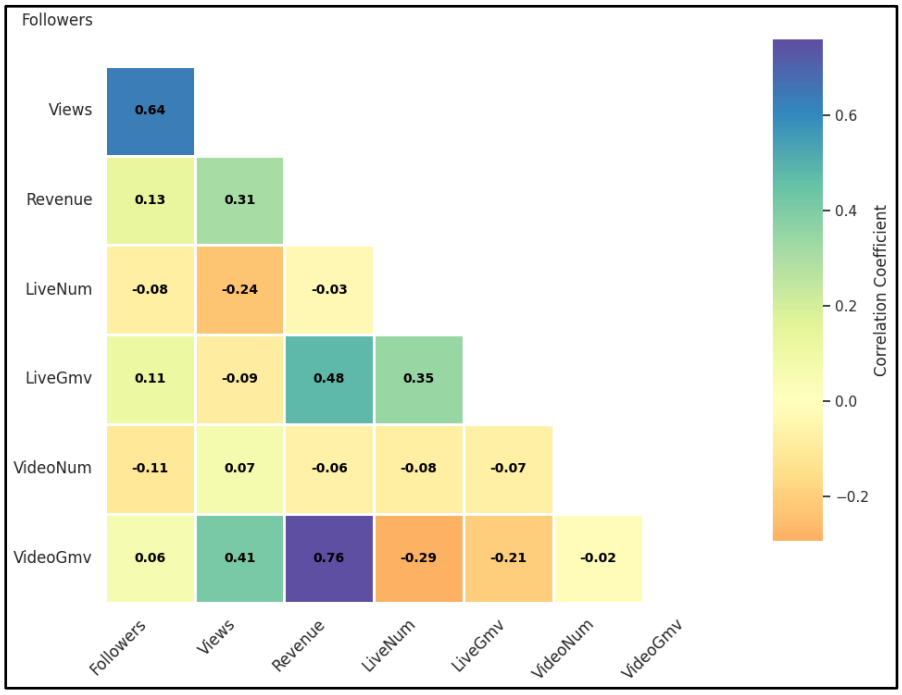


Figure 4.1 Correlation Matrix Of Revenue Data

Analysing the correlation matrix of revenue-related factors for TikTok influencers in the beauty and personal care sector reveals important insights that can help improve their performance. There is a strong positive relationship between revenue and sales from videos, as shown by a high correlation of 0.76 between revenue and video sales (VideoGmv). This means that influencers who generate more product sales through their video content tend to earn higher overall revenue. Therefore, focusing on creating engaging and persuasive video content that effectively promotes products can significantly boost earnings. Additionally, there is a moderate positive correlation of

0.48 between revenue and sales from live sessions (LiveGmv), indicating that live streaming can also contribute to higher income. Enhancing live session strategies such as by offering exclusive deals, interactive product demonstrations, or engaging directly with viewers can encourage more purchases during these sessions.

Interestingly, the number of followers has only a weak positive correlation with revenue (0.13), suggesting that simply having a large follower count does not guarantee higher earnings. However, there is a moderate positive correlation between followers and views (0.64), showing that a larger audience can lead to more content views. This implies that influencers should not just aim to grow their follower numbers but also focus on converting existing followers into customers by providing personalized content and targeted promotions. The moderate positive correlation between views and revenue (0.31) further supports the idea that increasing content visibility can help boost income, but it needs to be coupled with strategies that encourage viewers to make purchases.

The data also shows that the number of videos produced has almost no correlation with revenue (-0.06), and the number of live sessions has a minimal negative correlation with revenue (-0.03). This indicates that producing more content does not necessarily lead to higher earnings. Instead, it is the quality and effectiveness of the content that matter more. Influencers should prioritize creating high-quality videos and live sessions that resonate with their audience and showcase products effectively. There is a moderate positive correlation between the number of live sessions and sales from those sessions (0.35), suggesting that hosting regular live streams can increase sales during these events. However, it is important to balance the time and effort spent on live sessions and videos to ensure that one does not draw away from the other.

Lastly, there is a moderate positive correlation between views and video sales (0.41), highlighting that videos with more views tend to generate more product sales. To capitalize on this, influencers can work on making their videos more appealing and shareable. This could involve using eye-catching thumbnails and engaging storytelling to attract and retain viewers. By focusing on these areas, influencers can enhance their ability to turn viewers into customers. Overall, the insights from the correlation matrix suggest that influencers should concentrate on creating high-quality, engaging content that not only attracts a larger audience but also effectively encourages them to make purchases. By doing so, they can improve their revenue generation on TikTok and achieve greater success in the beauty and personal care sector.

Lastly, the Top 10 influencer analysis is conducted to identify the influencers that excel in generating revenue. The analysis involves ranking the influencers based on their revenue, revenue per follower, revenue per video, and revenue per live, while selecting the top 10 performer for each metric. This analysis will provide insights on these top influencer characteristics on the factors that contributes to their success. Due to the created variable revenue per follower, revenue per video, and revenue per live during the data preparation phase, in this section, the process is to only select the top 10 based on the variables created. The python code to select the top 10 influencers is shown below and figure 4.2 illustrates its result.

```
# Assigning the top 10 influencers to a variable
top_revenue_performers = revenue_analytics_df.sort_values(by='Revenue', ascending=False).head(10)

# Displaying the top 10 influencers
top_revenue_performers[['Nickname', 'Revenue', 'RevenuePerFollower', 'RevenuePerVideo',
'RevenuePerLive']]
```

The python code to select and display the top 10 influencers

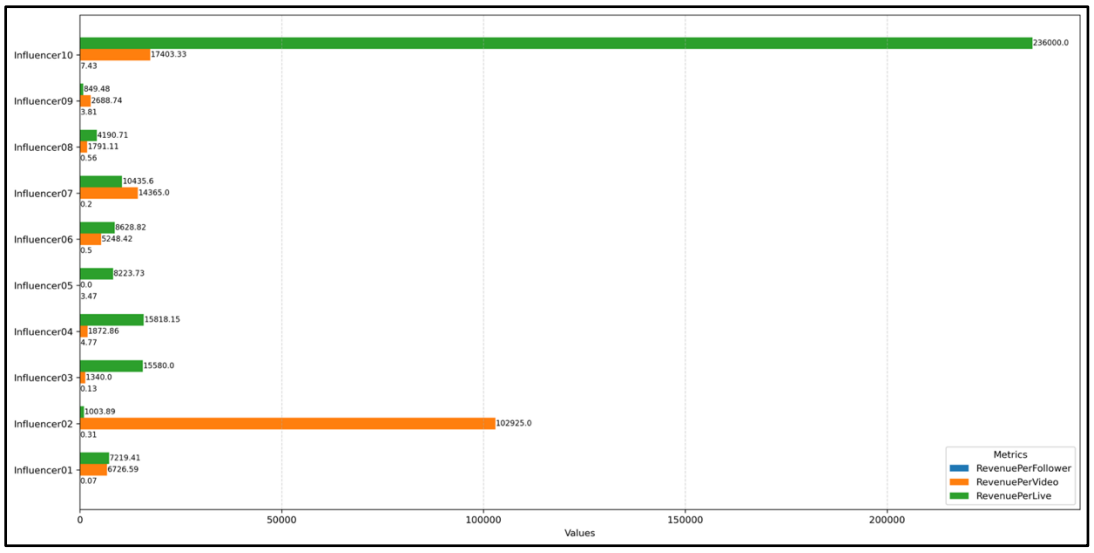


Figure 4.2 The Top 10 Influencers Based On Their Revenue

Based on Figure 4.2 of the top 10 TikTok influencers in the beauty and personal care sector ranked on their revenue reveals insightful patterns that can guide others toward greater success on the platform. Influencers like Influencer10, who leads with a revenue of RM1,540,000 and an impressive revenue per follower of RM7.43, demonstrate that high earnings are not solely dependent on having a vast follower base

but rather on effectively converting followers into customers. This influencer's substantial revenue per live session of RM17,403.33 suggests that engaging and interactive live streams significantly contribute to overall income. Similarly, Influencer02 showcases an exceptional revenue per video of RM102,925.00, despite a modest revenue per follower of RM0.31, indicating that producing impactful and high-quality video content can lead to significant earnings even without a large audience.

The data also highlights that influencers who diversify their content and revenue streams tend to achieve higher overall earnings. For instance, Incluencer07 generates considerable revenue from both videos and live sessions, emphasizing the importance of not relying solely on one type of content for monetization. On the other hand, influencers like Incluencer05, who has a revenue per video of RM0.00 but earns from live sessions, may be missing out on potential income by not monetizing all available content formats. This suggests that exploring and optimizing different avenues, such as adding product placements or partnering with brands in videos, can enhance overall revenue.

Moreover, some influencers with high total earnings, like Influencer01, have a low revenue per follower (RM0.07), indicating an unexplored potential within their existing audience. This implies that focusing on strategies to better convert followers into customers such as personalized content, exclusive promotions, and stronger calls to action can significantly boost revenue without necessarily increasing the follower count. Influencers like Influencer04, who has a high revenue per live session of RM15,818.15, illustrate the effectiveness of engaging live sessions in driving sales. By enhancing interactive elements, answering audience questions in real-time, and demonstrating products during live streams, influencers can encourage immediate purchases and increase their earnings.

In conclusion, the analysis of the top-earning influencers underscores that substantial revenue on TikTok is achieved not just through large follower counts but through effective engagement and monetization strategies. Prioritizing the creation of high-quality, impactful content that resonates with the audience, diversifying content types and income streams, and focusing on converting followers into customers are key practices that lead to greater financial success. By adopting these approaches, other influencers in the beauty and personal care sector can enhance their reach, engagement, and earnings on the platform, ultimately achieving better results in their social media endeavours.

4.4 Metadata Analytics

On the other hand, the metadata analytics also will start with calculating the summary statistics for variables such as "*playCount*", "*duration*", "*likesCount*", "*commentCount*", "*shareCount*", "*savesCount*", "*hashtagCount*", and "*engagement_rate*". The summary statistic involves calculating the mean, median, standard deviation, minimum, and maximum which could provide a clear picture of the engagement levels in the dataset. From this analysis, it could help in understanding the typical engagement a posts/video receives and the extent of variation across different posts/videos. The python code used to calculate all the statistics is displayed below and the results of the statistics is showed in the Table 4.4.

Table 4.4
Summary Statistic Of Video Metadata

Parameter	Mean	Media	Mode	Min	Max	Standard Deviation
play Count	355217.72	57150.00	1200000.00	566.00	27000000.00	1141025.22
duration	58.63	35.00	0.00	0.00	500.00	64.86
likes Count	16107.93	764.50	84.00	0.00	879400.00	48892.68
comment Count	172.82	24.00	1.00	0.00	32200.00	924.76
share Count	500.02	30.00	3.00	0.00	222800.00	5343.07
saves Count	968.17	84.00	3.00	0.00	64100.00	3261.37
hashtag Count	2.72	0.00	0.00	0.00	22.00	4.43
Engagement rate	0.03	0.02	0.03	0.00	0.27	0.03

```
# List of the metrics
metrics = ['playCount', 'duration', 'likesCount', 'commentCount', 'shareCount', 'savesCount',
'hashtagCount', 'engagement_rate']

# Create an empty dictionary to store the results
statistics = {}

for metric in metrics:
    mean_val = round(metadata_analytics_df[metric].mean(), 2)
    median_val = metadata_analytics_df[metric].median()
    mode_val = metadata_analytics_df[metric].mode()[0]  # Taking the first mode in case of multiple
modes
    min_val = metadata_analytics_df[metric].min()
    max_val = metadata_analytics_df[metric].max()
    std_val = round(metadata_analytics_df[metric].std(), 2)

    # Store the results in the dictionary
    statistics[metric] = {
        'Mean': mean_val,
        'Median': median_val,
        'Mode': mode_val,
        'Min': min_val,
        'Max': max_val,
        'Standard Deviation': std_val
    }

# Convert dictionary to DataFrame for better visualization
statistics_df = pd.DataFrame(statistics)

# Format the statistics to 2 decimal places
statistics_df[metrics] = statistics_df[metrics].applymap('{:.2f}'.format)

# Transpose the DataFrame to have metrics as rows and stats as columns
statistics_df = statistics_df.T

# Print the results
statistics_df
```

Python code to calculate the summary statistics for video metadata

Analysing the metadata of 2,016 TikTok videos from 20 influencers in the beauty and personal care sector reveals significant opportunities to enhance content performance on the platform. The average play count stands at approximately 355,218 views, yet the median is much lower at 57,150 views, indicating that while a few videos achieve viral status with up to 27 million views, the majority obtain only moderate attention. This imbalance value suggests that influencers could benefit from studying their most successful content to identify elements that contribute to higher engagement,

such as compelling storytelling or trending topics, and replicating those strategies in future videos. The wide range in video duration, from instant clips to lengthy 500-second recordings, with an average of about 59 seconds and a median of 35 seconds, highlights the importance of optimizing content length. Focusing on creating concise and impactful videos between 15 and 60 seconds may enhance viewer retention and increase the likelihood of shares, as shorter videos tend to hold viewers' attention more effectively on TikTok.

Engagement metrics such as likes, comments, shares, and saves also exhibit significant gaps between their averages and medians, with the mean likes count at over 16,000 while the median is only around 765. This indicates that most videos receive fewer interactions, and the high averages are skewed by a small number of exceptionally popular posts. To address this, influencers should aim to create content that encourages audience interaction by including clear calls to action, posing questions, or incorporating interactive elements like polls and challenges. Interestingly, although the average hashtag count is 2.72, both the median and mode are zero, suggesting that many videos do not utilize hashtags. Given that hashtags play a crucial role in content discoverability on TikTok, influencers should incorporate relevant and trending hashtags to increase their reach and attract new viewers searching for specific topics.

The average engagement rate of 3%, with a median of 2%, points to potential growth in how audiences interact with content. Enhancing engagement rates can be achieved by crafting content that resonates emotionally with viewers, offers value through tutorials or tips, or leverages trending topics to make videos more relatable and shareable. Additionally, the average saves count of about 968, contrasted with a median of only 84, indicates that while some content is deemed valuable enough for viewers to revisit, many videos do not achieve this level of impact. By creating informative or inspirational content that provides lasting value, influencers can encourage viewers to save videos, signalling stronger engagement and interest.

Overall the summary statistics of metadata underscores the importance of producing high-quality, engaging content that not only captures viewers' attention but also encourages interaction and sharing. By focusing on optimal video lengths, effectively using hashtags, incorporating interactive elements, and aligning content with audience interests and trends, influencers in the beauty and personal care sector can enhance their visibility, foster deeper connections with their audience, and ultimately achieve greater success on TikTok.

Furthermore, the analysis continues to examine the correlation matrix for metadata-related variables. Based on the Pearson correlation coefficients between variables like “playCount”, “duration”, “likesCount”, “commentCount”, “shareCount”, “savesCount”, “hashtagCount”, and “engagement_rate”, the key relationships between these variables could be identified and how different type of engagement are connected also could be understand. To calculate the correlation, the Pandas library is utilized due to its convenient compared to another library such as NumPy and SciPy. Below shows the python codes to utilize the Pandas library to calculate the Pearson correlation coefficients and Figure 4.3 shows the calculated correlation matrix.

```
# Calculate the correlation between duration and other metrics
correlations = metadata_analytics_df[['duration', 'likesCount', 'commentCount', 'shareCount',
'savesCount', 'hashtagCount', 'engagement_rate']].corr()

# Display correlation coefficients
correlations
```

Python code to calculate the correlation matrix

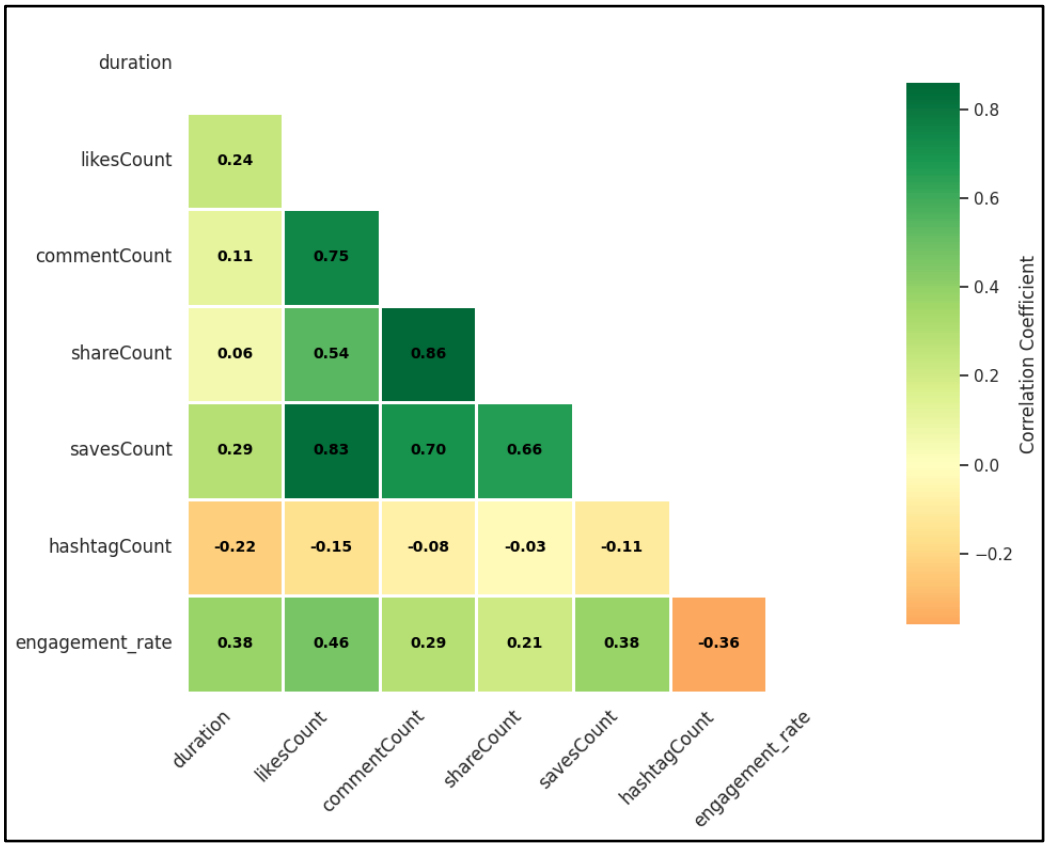


Figure 4.3 Correlation Matrix For Video Metadata

Analysing the correlation matrix for metadata related-variables on figure 4.3, reveals several key relationships that can inform strategies to enhance content performance on the platform. The data shows a moderate positive correlation between video duration and engagement rate (0.3788), suggesting that slightly longer videos tend to engage viewers more effectively. This implies that influencers might benefit from creating content that provides more in-depth information or storytelling, capturing the audience's attention for a longer period. Additionally, there is a strong positive correlation between likes count and saves count (0.8258), indicating that videos that receive more likes are also more likely to be saved by viewers. This relationship suggests that creating valuable or memorable content encourages viewers to both like and save the videos for future reference.

The likes count also has a strong positive correlation with comment count (0.7453) and share count (0.5397), highlighting that when viewers appreciate content enough to like it, they are also more inclined to comment and share. This underscores the importance of producing engaging content that resonates with the audience on multiple levels. Interestingly, there is a negative correlation between hashtag count and engagement rate (-0.3598), as well as between hashtag count and duration (-0.2228). This suggests that using too many hashtags may detract from viewer engagement and could be associated with shorter video lengths. Influencers might consider limiting the number of hashtags to those most relevant to their content to avoid overwhelming or distracting viewers.

The strong positive correlation between comment count and share count (0.8584) indicates that videos prompting viewers to comment are also more likely to be shared. Encouraging audience interaction through questions or calls to action can therefore amplify both comments and shares, extending the content's reachability. Similarly, the positive correlation between saves count and engagement rate (0.3788) implies that when viewers save videos, it reflects a higher overall engagement with the content. Creating informative or inspirational videos that provide lasting value can encourage viewers to save them, enhancing engagement metrics.

Furthermore, the data shows that while likes count has a moderate positive correlation with engagement rate (0.4647), comment count and share count have weaker positive correlations with engagement rate (0.2941 and 0.2069, respectively). This indicates that likes are a more significant driver of engagement rate compared to comments and shares. Influencers should therefore prioritize strategies that encourage

viewers to like their videos, such as by creating relatable content or using compelling visuals.

Overall, the analysis suggests that influencers can improve their TikTok performance by focusing on creating slightly longer videos that provide value and encourage viewers to like, save, comment, and share. Limiting the use of hashtags to the most relevant ones may enhance engagement rates, as excessive hashtags could negatively impact viewer interaction. By understanding these relationships and tailoring their content strategies accordingly, influencers can foster deeper connections with their audience, increase engagement metrics, and achieve greater success on the platform.

Based on the correlation matrix, it is shown that the number of hashtags is negatively correlate with engagement rate. Thus, the analysis moves forward to delve deeper into the hashtag analysis. This analysis calculates the average engagement rate for posts with different number of hashtags to further see the relationship between the number of hashtags and the engagement rate. For instance, what are the number of hashtags that could contributes to a higher engagement rate. By achieving this analysis, it could provide the insights into the effectiveness of hashtag strategies and help optimize future content for better engagement. To do this analysis, Pandas, python library is used to grouped variables like “likesCount”, “commentCount”, “shareCount”, “playCount”, “savesCount”, “engagement_rate” with the “hashtagCount”. The result is then sorted based on the number of hashtags. This analysis is represented graphically to show the trends of the number of hashtags and the average engagement rate. Python code to conduct this analysis is shown below and Figure 4.4 shows the results of the analysis.

```
# Calculate the average engagement metrics based on the hashtag count
hashtag_count_engagement = metadata_analytics_df.groupby('hashtagCount')[['likesCount',
'commentCount', 'shareCount', 'playCount', 'savesCount', 'engagement_rate']].mean().reset_index()

# Formatting the digits
columns_to_format = ['likesCount', 'commentCount', 'shareCount', 'savesCount', 'playCount']  # List
of columns to format

for col in columns_to_format:
    hashtag_count_engagement[col] = hashtag_count_engagement[col].apply(lambda x: f"{float(x):,.2f}")

# Display the aggregated metrics
hashtag_count_engagement.sort_values(by='hashtagCount', ascending=True)

# Plot the average engagement rate based on hashtag count
plt.figure(figsize=(10, 6))
sns.barplot(x='hashtagCount', y='engagement_rate', data=hashtag_count_engagement, palette='viridis')
plt.title('Impact of Hashtag Count on Engagement Rate')
plt.xlabel('Number of Hashtags')
plt.ylabel('Average Engagement Rate')
plt.show()
```

The python code to conduct the hashtag count analysis

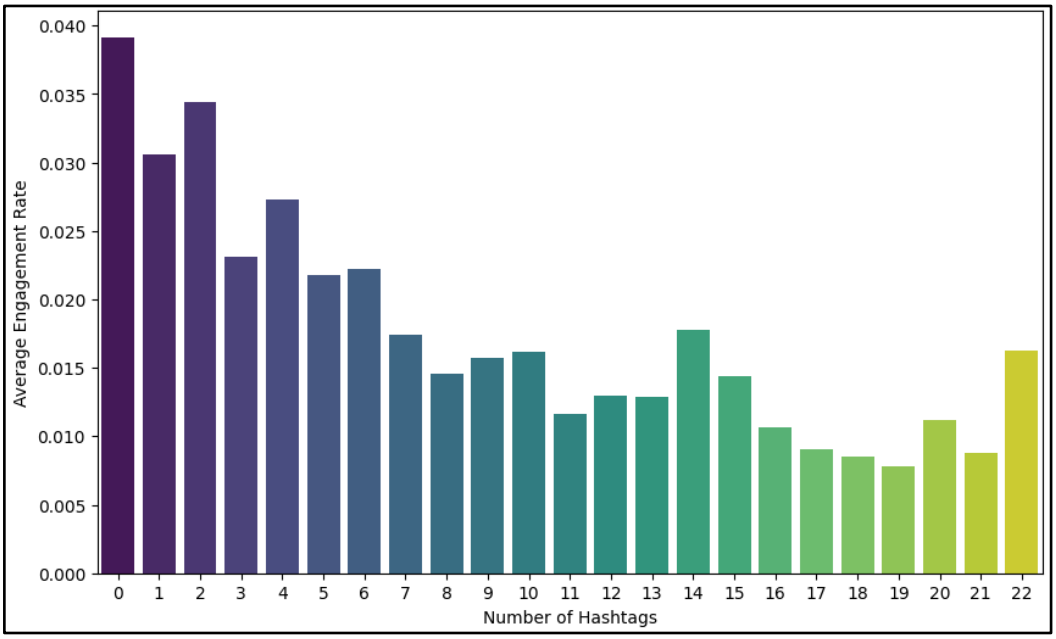


Figure 4.4 The Number Of Hashtags And The Average Engagement Rate

Analysing the relationship between the number of hashtags used and the corresponding engagement rates shows a compelling inverse correlation that offers valuable guidance for influencers aiming to enhance their content performance. The data clearly indicates that as the number of hashtags increases, the engagement rate consistently decreases. Specifically, videos that lack of hashtags possess the highest average engagement rate of approximately 3.92%, while those incorporating just one or two hashtags maintain a commendable engagement rate around 3%. In contrast, content

featuring more than three hashtags experiences a notable decline in audience interaction, with engagement rates shrinking below 2% and reaching as low as 0.78% for videos with nineteen hashtags.

This trend suggests that excessive use of hashtags may inadvertently dilute the video's impact, potentially overwhelming viewers or distracting attention away from the core message. Instead of enhancing discoverability, an abundance of hashtags might make the content appear cluttered or inauthentic, thereby diminishing viewer engagement. Moreover, the data reveals that videos with fewer hashtags not only achieve higher engagement rates but also gain a greater averages in likes, comments, shares, and saves. For instance, videos without hashtags average 21,905 likes, significantly outperforming those with numerous hashtags.

These insights underscore the importance of strategic hashtag utilization, where quality supersedes quantity. Influencers are encouraged to limit their hashtag usage to no more than two per video, selecting only the most relevant and trending tags that align closely with their content. This approach ensures that hashtags serve their intended purpose of enhancing discoverability without detracting from the viewer's experience. By focusing on the substance of the content and maintaining a clean presentation, creators can foster a stronger connection with their audience. Furthermore, the analysis highlights that increasing the number of hashtags does not proportionally boost other engagement metrics or play counts. Videos with fewer hashtags actually enjoy higher average play counts, suggesting that viewers may be more inclined to watch and engage with content that appears more authentic and less cluttered. This reinforces the notion that overusing hashtags does not translate into increased visibility or popularity.

Through these findings, influencers should consider adopting a minimalist approach to hashtag usage. By prioritizing high-quality content and resisting the temptation to overload videos with hashtags, they can enhance viewer engagement and satisfaction. Experimenting with different hashtag counts and closely monitoring engagement metrics can help creators identify the optimal balance that resonates with their specific audience. Ultimately, focusing on meaningful connections with viewers through compelling content and thoughtful hashtag selection can lead to greater success on the TikTok platform. Finally, in this research, the temporal analysis is performed to see how engagement metrics changes over time. To identify the trends and seasonal patterns, the data is grouped by the "dateCreated" variables and the average values for "likesCount", "playCount", "shareCount", and "commentCount" is calculated. This

analysis will help in understanding how engagement evolves and what are the factors influences the changes. The steps to execute this analysis is by making sure that the “dateCreated” variable is in date data type and the average number of metrics “likesCount”, “playCount”, “shareCount”, and “commentCount” is grouped by the date. The results are then visualised in graph. Below is the python code to check the data type and calculate the average values for each metric while figure 4.6, 4.7, 4.8, and 4.9 shows the visualisations for each of the engagement metrics.

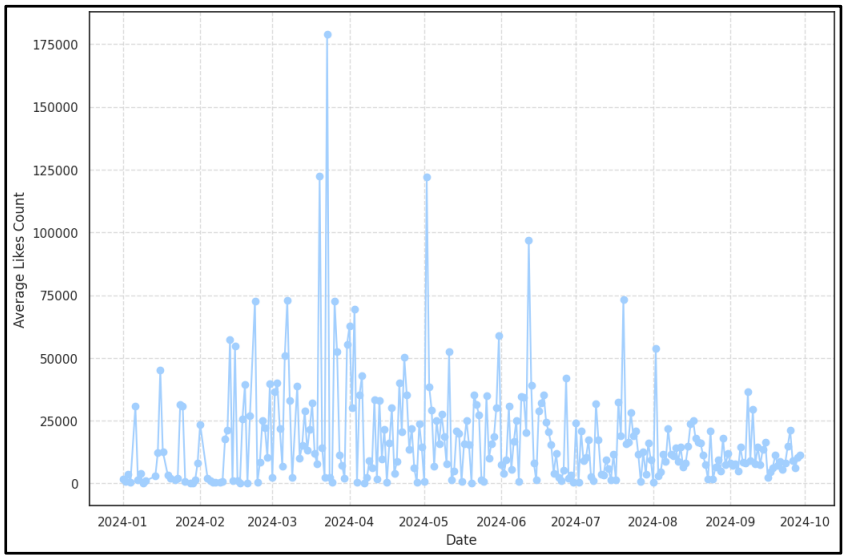


Figure 4.5 The Temporal Analysis For Likes Count

```
# Ensure 'dateCreated' is in datetime format
metadata_analytics_df['dateCreated'] = pd.to_datetime(metadata_analytics_df['dateCreated'])

# Group by date and calculate mean engagement metrics
temporal_analysis = metadata_analytics_df.groupby(metadata_analytics_df['dateCreated'].dt.date).agg({
    “likesCount”: “mean”,
    “playCount”: “mean”,
    “shareCount”: “mean”,
    “commentCount”: “mean”
})

# Plotting the results
plt.figure(figsize=(14, 7))

# Plot likesCount over time
plt.subplot(2, 2, 1)
plt.plot(temporal_analysis.index, temporal_analysis['likesCount'], marker='o')
plt.title('Average Likes Count Over Time')
plt.xlabel('Date')
plt.ylabel('Average Likes Count')

# Plot playCount over time
plt.subplot(2, 2, 2)
plt.plot(temporal_analysis.index, temporal_analysis['playCount'], marker='o')
plt.title('Average Play Count Over Time')
plt.xlabel('Date')
plt.ylabel('Average Play Count')

# Plot shareCount over time
plt.subplot(2, 2, 3)
plt.plot(temporal_analysis.index, temporal_analysis['shareCount'], marker='o')
plt.title('Average Share Count Over Time')
plt.xlabel('Date')
plt.ylabel('Average Share Count')

# Plot commentCount over time
plt.subplot(2, 2, 4)
plt.plot(temporal_analysis.index, temporal_analysis['commentCount'], marker='o')
plt.title('Average Comment Count Over Time')
plt.xlabel('Date')
plt.ylabel('Average Comment Count')

plt.tight_layout()
plt.show()
```

The python code to conduct the temporal analysis

The temporal analysis of engagement metrics for TikTok influencers in the beauty and personal care sector starts from January to September 2024. The metrics

examined include likes, play counts, share counts, and comment counts, each offering a window into how audiences are engaging with content over time.

Based on Figure 4.6, the temporal analysis of likes count, there is a notable fluctuation throughout the months, with certain dates exhibiting significant spikes. For instance, on February 13th, there is a remarkable increase to an average of 57,226 likes, suggesting that content posted around this time resonated strongly with the audience. Similarly, March 23rd shows an even more substantial surge to approximately 178,911 likes, indicating highly engaging content or possibly the effect of a viral trend or challenge during that period. These peaks contrast with much lower averages on dates like February 20th, where likes dropped to around 13, pointing to less engaging content or reduced audience activity. The fluctuation in likes suggests that timing and content relevance play crucial roles in capturing audience appreciation. Influencers should analyse the content posted during high-like periods to identify successful elements, such as themes, presentation styles, or topics, and aim to replicate these in future posts. Additionally, aligning content with seasonal trends or events that elicit higher audience interest can boost likes.

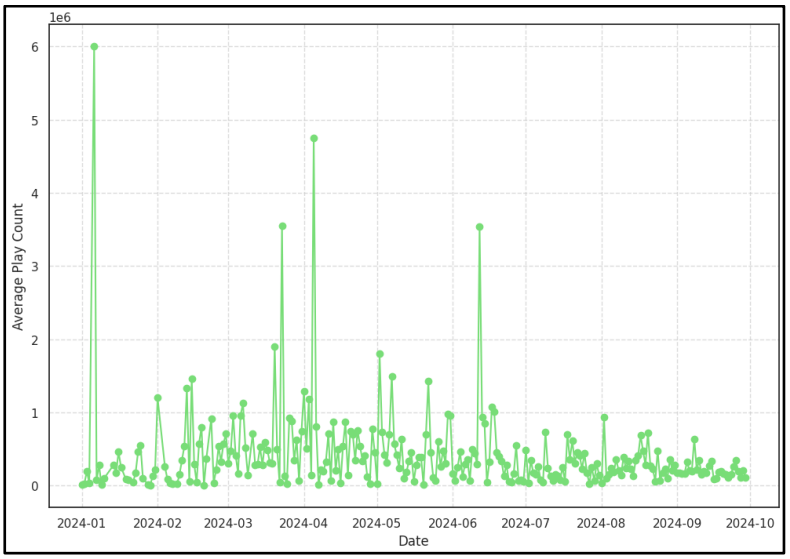


Figure 4.6 The Temporal Analysis For Play Counts

Moving to play counts which illustrated in Figure 4.7, the data indicates that viewer engagement in terms of video views also experiences significant variability. High play counts are observed on dates such as January 6th, with an impressive average of 6,000,000 plays, and on March 20th, reaching up to 1,900,000 plays. These surges

may be attributed to the release of particularly captivating content, collaborations with other popular influencers, or participation in trending topics that attract widespread attention. In contrast, lower play counts on dates like February 20th, with averages as low as 1,479 plays, suggest less effective content or decreased platform activity. To maximize play counts, influencers should focus on creating compelling content that aligns with current trends and encourages viewers to watch. Collaborations and strategic posting times, when audience activity is highest, can also enhance visibility and attract more views.

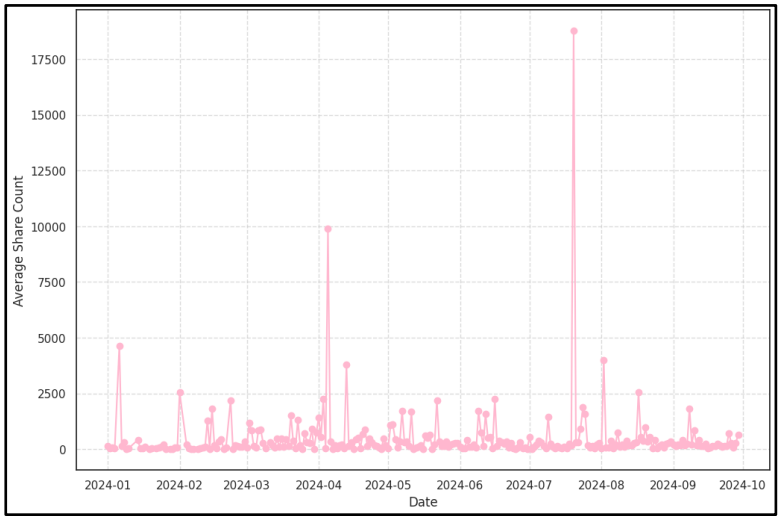


Figure 4.7 The Temporal Analysis For Share Counts

Regarding share counts (as illustrated in Figure 4.8), which reflect how often viewers are compelled to share content with others, there are notable peaks and troughs throughout the data. Significant increases in share counts occur on dates like February 15th, with an average of 1,826 shares, and May 2nd, with a substantial 1,071 shares. These spikes indicate that the content was not only engaging but also deemed valuable enough for viewers to share within their networks. On the other hand, minimal share counts on dates such as January 28th, where shares dropped to zero, highlight content that may not have resonated sufficiently to prompt sharing. To enhance shareability, influencers should create content that evokes strong emotional responses, provides valuable information, or aligns with viral trends, motivating viewers to share. Incorporating calls to action that encourage sharing can also be effective.

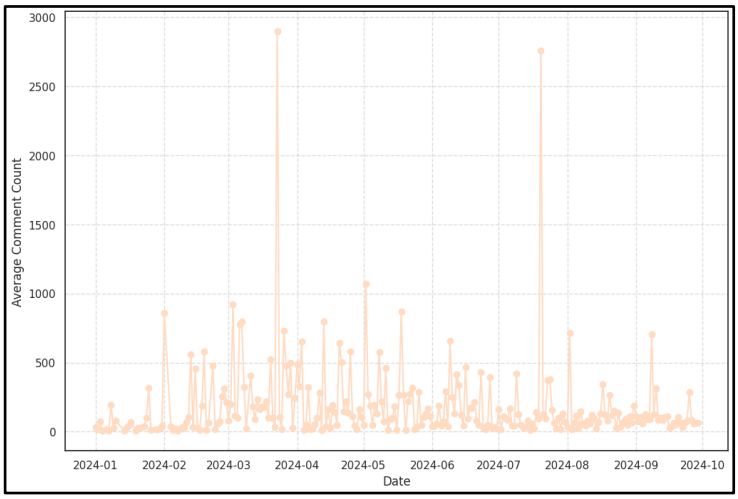


Figure 4.8 The Temporal Analysis For Comment Counts

Lastly, examining the temporal analysis of comment counts based on figure 4.9, which offer insight into the level of audience interaction and engagement with the content, reveals patterns similar to the other metrics. High comment counts are observed on dates like March 23rd, with an exceptional average of 2,898 comments, suggesting highly engaging content that sparked conversations among viewers. Conversely, lower comment counts on dates like January 4th, with an average of 9.5 comments, indicate less interactive content. To increase comments, influencers should craft content that invites discussion, such as posing questions, encouraging opinions, or addressing relatable topics that prompt viewers to share their thoughts. Engaging with commenters by responding can further stimulate interaction and build a sense of community.

Overall, the temporal analysis of engagement metrics underscores the importance of strategic content creation and timing, particularly in relation to key events that occur in Malaysia. Influencers should pay close attention to the types of content that generate higher likes, plays, shares, and comments, while also aligning their posts with cultural and festive periods. For instance, engagement spikes around Chinese New Year (February 10, 2024), Ramadan and Hari Raya Aidilfitri (March-April 2024), and Merdeka Day (August 31, 2024) suggest that content resonating with these events such as festive makeup tutorials, traditional attire inspiration, or patriotic-themed beauty looks, can naturally boost engagement. Similarly, the Kaamatan and Gawai festivals (May and June) and Hari Malaysia (September 16) present opportunities for content that celebrates local culture and beauty practices, attracting increased shares and comments. By tailoring their content strategies to replicate successful elements identified during these peak times and fostering interactive experiences tied to cultural

events, influencers can enhance their overall audience engagement and achieve greater success on the platform.

4.5 Conclusions

In conclusion, this chapter analyzed TikTok data within the beauty and personal care sector using descriptive analytics techniques, including exploratory data analysis, revenue analytics, and metadata analytics. The analysis reveals several critical patterns and relationships, such as the influence of content quality on revenue generation and the impact of engagement strategies on audience interaction. These analyses collectively addressed Research Objective 2 (RO2) by identifying the patterns, trends, and engagement factors that influence audience behavior on the platform. Table 4.4 below summarizes the key insights from this chapter.

Table 4.5
Summary of Insights

Analytics	Insights
Exploratory Data Analysis	• The data reveals significant variability in follower counts and engagement metrics, suggesting a highly uneven distribution of success among influencers. • Average engagement metrics highlight opportunities for influencers to optimize content strategies and better cater to their audience's preferences. • High averages in key metrics such as play counts and likes are often driven by outliers, showing the importance of understanding top-performing content.
Revenue Analytics	• Influencers with diversified revenue streams such as video and live sessions, consistently outperform others, emphasizing the need for a multi-channel monetization strategy. • A weak correlation between follower counts and revenue shows that monetization depends more on audience engagement and effective conversion tactics than on vast reach. • High variation in revenue per follower indicates opportunities to implement personalized engagement tactics and exclusive promotions to maximize earnings from existing audiences.
Metadata Analytics	• Engagement rates decline as hashtag counts increase beyond two, underscoring the importance of using fewer, highly relevant hashtags to avoid clutter and maximize discoverability. • Likes, shares, and saves are strongly correlated, demonstrating that content that resonates emotionally or

	provides lasting value tends to drive higher audience interaction. • Videos of slightly longer durations (around 35–60 seconds) perform better in engagement metrics, pointing to the value of well-crafted and in-depth content. • Temporal analysis shows that aligning content with cultural or seasonal events leads to significant spikes in engagement, offering an opportunity for better content timing.

While each analytical phase provides distinct findings, their true value lies in their interconnection, forming a holistic framework for actionable business intelligence. The Exploratory Data Analysis (EDA) first established the "performance gap", revealing that a small percentage of influencers generate the majority of revenue and engagement. The Revenue Analytics then deconstructed this gap, proving that follower count alone is a poor predictor of financial success (weak correlation). Instead, it identified that active content strategies like sales generated through videos and live sessions are the primary drivers of revenue. This finding directs the focus from "vanity metrics" (followers) to "conversion metrics". This is where Metadata Analytics becomes critical, as it provides the tactical "blueprint" to achieve that conversion. By identifying that lower hashtag counts (0-2) and specific video durations (35-60 seconds) correlate with higher engagement and metadata analytics offers the specific content optimizations needed to drive the sales identified in the revenue analysis. Finally, the Temporal Analysis acts as the delivery mechanism, ensuring that this optimized content is deployed during peak cultural windows (like festivals) to maximize reach. Collectively, these analytics reinforce each other, Revenue Analytics defines the goal (conversion over followers), Metadata Analytics provides the method (content optimization), and Temporal Analysis dictates the timing. Together, they generate the actionable insight that successful TikTok strategies must pivot from passive follower growth to active, quality-driven, and culturally timed content creation.

While the descriptive analytics in this chapter successfully quantified what is happening through showing that engagement drives revenue, they do not explain why audiences engage or how they perceive the content emotionally. Mere numbers (likes and views) cannot capture the nuances of consumer trust, satisfaction, or dissatisfaction. To address this gap and fulfil Research Objective 3 (RO3), the research must move beyond quantitative metrics to qualitative understanding. Consequently, Chapter 5 will apply advanced Natural Language Processing (NLP) techniques, specifically the GPT-

4o model, to conduct a sentiment analysis on the collected comments. This will allow for the extraction of deeper business intelligence by categorizing consumer perceptions into Positive, Neutral, and Negative sentiments, providing the “why” behind the engagement trends observed in this chapter.

CHAPTER 5
SENTIMENT ANALYSIS

5.1 Introduction

This chapter presents the application of sentiment analysis as an advanced modeling technique to extract business intelligence insights from TikTok comments, with a particular focus on public perceptions toward influencers in the Beauty and Personal Care category. The analysis is conducted using the GPT-4o model to classify comments into Positive, Neutral, and Negative sentiments. The dataset used in this process has undergone cleaning and standardization, as detailed in the previous chapter. Additionally, this research references an existing research paper as a benchmark, providing a structured approach for implementing GPT-based sentiment analysis on TikTok data.

5.2 Data Overview

This section offers a high-level summary of the dataset employed for GPT-based sentiment analysis. The data was sourced from TikTok's Beauty and Personal Care category, focusing on the top 20 influencers based on revenue generation. Video content was extracted using Apify's TikTok Profile Scraper API, covering posts from January 1, 2024, to August 28, 2024. Each influencer's profile yielded up to 300 recent videos, ultimately totalling 3,912 videos. Subsequently, up to 100 comments per video were retrieved using the TikTok Comments Scraper API, resulting in 34,597 comments. Table 5.1 shows a summary of the data collection details.

Table 5.1
Summary Of Data Collection

Parameter	Details
Exploratory Data Analysis	Top 20 influencers with the highest total revenue in the Beauty and Personal Care category
Revenue Analytics	Up to 300 videos per influencer (actual count varies based on content production)
Metadata Analytics	Up to 100 most recent comments per video, ranked by engagement

Data collection date	28 August 2024
Period of the collected videos	1 January 2024 – 28 August 2024
Total number of videos collected	3,912
Total number of comments collected	34,597
Number of comments after grouping by video	2,016

A brief preprocessing pipeline was then applied to the collected comments to enhance data quality and consistency, which involved normalizing slang, expanding local abbreviations, and removing irrelevant or empty entries. These processes are discussed in detail in Chapter 3. In addition to preparing the comments for sentiment analysis, the emojis were converted to textual representations (e.g., ":face_with_tears_of_joy:") to capture emotional context. These steps ensure that informal or symbolic expressions are recognized accurately during sentiment classification. Once the dataset is cleaned and standardized, it serves as the basis for GPT-based sentiment analysis in the subsequent sections.

5.3 GPT-Based Approach

Sentiment analysis has evolved from rule-based and lexicon-driven approaches to advanced deep learning and transformer-based techniques (Gandhi et al., 2021; Pang et al., 2002). Although earlier methods such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks improved performance on social media data (Gandhi et al., 2021), challenges like sarcasm, slang, and multilingual expressions persist (Hartmann et al., 2023; Ray & Chakrabarti, 2022). Transformer-based models, particularly BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), addressed some of these complexities by leveraging self-attention mechanisms and large-scale pretraining (Nguyen et al., 2020). However, the informal and emoji-rich nature of TikTok comments still presents issues of brevity and code-switching, which can diminish the effectiveness of standard transformer-based methods (Cheng & Li, 2024; B. Cho, 2024). Building on these developments, GPT-based sentiment analysis has gained traction for its capacity to interpret short-form, highly informal user-generated text without the need for extensive feature engineering (W. Zhang, Deng, et

al., 2023). In the Beauty and Personal Care domain on TikTok, user comments often combine local slang, emoji-based expressions, and multilingual elements (Ahmad Asmawi & Isawasan, 2024). The study by Kheiri & Karimi (2023) indicate that GPT can handle these linguistic nuances more effectively than classical ML or earlier transformer models. The GPT architecture's generative capacity and large-scale language modeling allow it to capture context and subtle cues such as sarcasm or emoji sentiment which significantly improves the classification consistency in short social media text.

In this research, GPT-4o was employed to classify TikTok comments into Positive, Neutral, or Negative sentiments. Following insights from the above-mentioned works, minimal prompt engineering was performed to define sentiment categories and clarify domain context. Default model parameters such as temperature and token limits remained unchanged, mirroring earlier findings that GPT's robust language model often requires minimal hyperparameter tuning for social media tasks (Elmitwalli & Mehegan, 2024; Mughal et al., 2024). The GPT-4o API served as the core classification engine, integrated within a Python environment. Data was loaded and processed using pandas and NumPy, while Matplotlib and Seaborn were employed for visualizing sentiment distributions and comment-level statistics. Text normalization steps (e.g., tokenization, removal of special symbols) were handled by Python-based NLP utilities. By aligning with best practices from prior GPT-based research (Gupta et al., 2023; Hutto & Gilbert, 2014), the data pipeline ensured a clean and consistent input for GPT-4o, thus maximizing the reliability of the sentiment outputs in the TikTok context.

5.4 Sentiment Analysis

5.4.1 Implementation Steps

Following the data preparation outlined in Chapter 3, the cleaned TikTok comments were input into a GPT-based sentiment classification pipeline. Figure 5.1 illustrates the workflow.

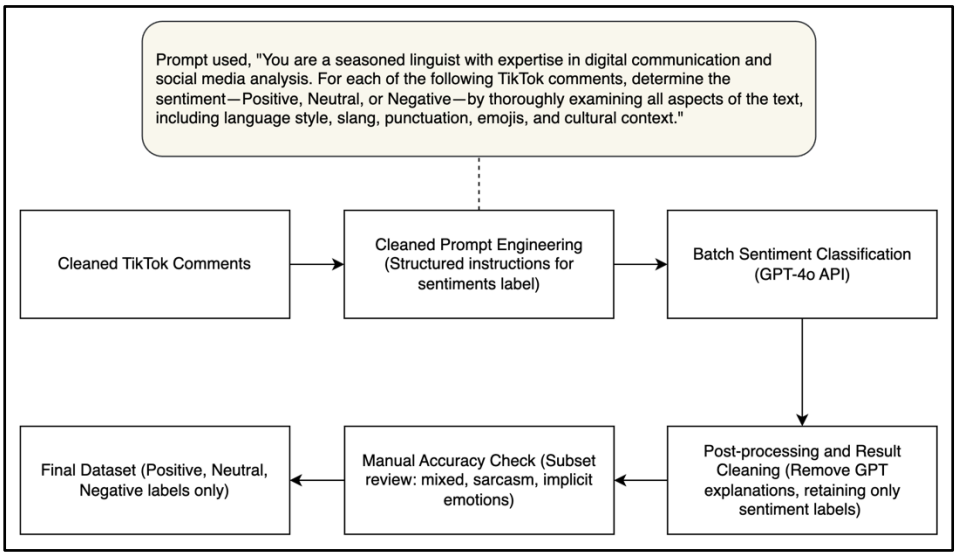


Figure 5.1 Overview Of GPT-Based Sentiment Classification Process

The GPT-based sentiment classification pipeline commenced with a minimal prompt design, specifying the three sentiment labels, Positive, Neutral, and Negative alongside a concise context statement to guide the model. This approach aligns with previous findings indicating that GPT can accurately interpret social media text with minimal hyperparameter tuning or complex prompt structures. After prompt engineering, the cleaned TikTok comments were batched and submitted to the GPT-4o model via an API, which then assigned each comment a sentiment label. Special text normalization rules, such as converting emojis into textual placeholders, were applied to ensure that GPT could accurately recognize and interpret emoji-driven emotional cues. A subset of comments underwent manual verification to evaluate GPT's accuracy and consistency. Observations from this manual review suggested that GPT's generative capacity allowed it to interpret informal language including slang and code-switching more effectively than traditional rule-based or machine learning approaches. Upon finalizing sentiment labels, the classification results were stored in a structured dataset, linking each comment to its corresponding influencer and video. This approach enabled deeper analyses of sentiment trends, frequently used words, and influencer-specific sentiment distributions, all of which are explored in the subsequent sections.

5.4.2 Results and Findings

5.4.2.1 Overall Sentiment Distribution

The bar chart in Figure 5.2 illustrates the sentiment breakdown among 2,016 TikTok comments. A total of 1,455 comments (72.18%) were labelled as Positive, 490 comments (24.28%) were Neutral, and 71 comments (3.52%) were Negative. This predominance of Positive sentiment may reflect the informal, entertainment-focused environment of TikTok, where users frequently leave supportive or enthusiastic feedback, especially within the Beauty and Personal Care category. The presence of 490 Neutral comments indicates that a significant portion of users is primarily seeking information or clarifications, rather than expressing strong approval or disapproval. These queries often revolve around product details or usage instructions, suggesting that potential buyers rely on TikTok as an informal research platform. Although the Negative sentiment constitutes only a small fraction (3.52%), the concerns raised such as dissatisfaction with product quality or misleading claims are crucial for influencers and brands aiming to maintain trust. While most comments are supportive, Neutral feedback should not be overlooked. Providing comprehensive, clear product information could help convert Neutral comments into positive engagement or even sales, underscoring the commercial importance of addressing user queries promptly.

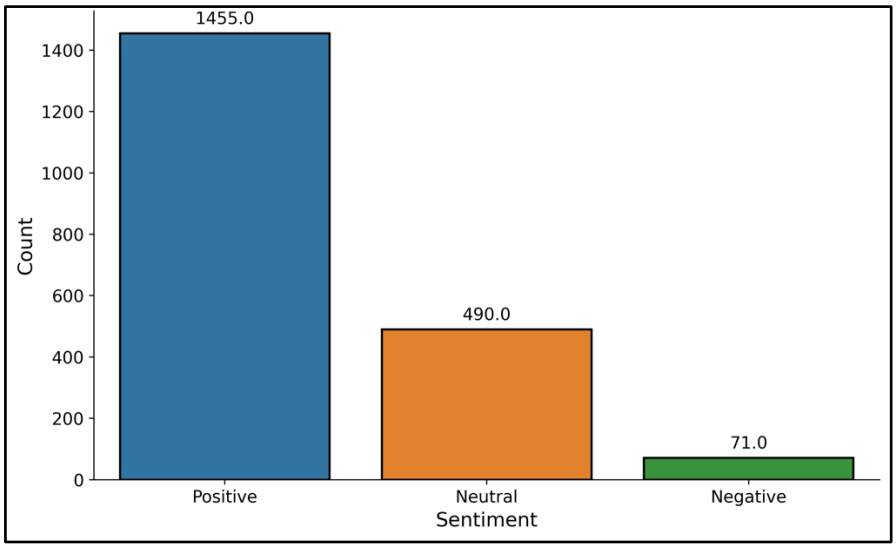


Figure 5.2 Distributions Of Sentiments

5.4.2.2 Top Words Across Sentiment Categories

The analysis identified the most frequent and contextually relevant words and emojis associated with each sentiment category. These terms highlight how consumers express different attitudes toward products and influencers in TikTok comments. By examining them, clearer patterns emerge in how Positive, Neutral, and Negative sentiments are communicated. In the Positive category, as shown in Table 5.2 below, comments are dominated by affectionate emojis such as "*smiling_face_with_hearts*", "*red_heart*", "*thumbs_up*", and "*face_with_tears_of_joy*", together with words of praise like "*cantik*", "*comel*", "*terbaik*", and "*sangat*". These terms reflect admiration and enthusiasm, often directed at the appearance of influencers or the perceived quality of products. Alongside expressions of appreciation, practical references such as "*botol*", "*bau*", "*wangi*", "*selesai*", "*beli*", "*solekan*", and "*pesanan*" show that positivity is tied not only to emotional reactions but also to consumer interest and purchase behavior. Cultural expressions like "*tahniah*", "*alhamdulillah*", "*semoga*", and "*selamat*" further amplify this positivity by embedding comments in broader social and religious contexts. Overall, Positive sentiment combines emotional reinforcement with concrete signals of product adoption and social approval.

Table 5.2
Positive Top Words

Top Words	Total Count
smiling_face_with_hearts	1880
red_heart	1617
face_with_tears_of_joy	1064
cantik	1026
thumbs_up	926
comel	743
terbaik	631
bau	621
wangi	593
beli	572
selesai	556
botol	538
tahniah	511

semoga	486
alhamdulillah	467
selamat	458
sangat	447
cantiknya	421
pesanan	397
solekan	382
tahan	368
best	357
baju	345
face_blowing_a_kiss	287
smiling_face_with_heart	279
rolling_on_the_floor_laughing	263
heart_suit	251

Neutral sentiment, as shown in Table 5.3, is shaped by inquiry-oriented language and evaluative terms. Words such as "*boleh*", "*mana*", "*kalua*", "*mahu*", "*sesuai*", "*tanya*", and "*adakah*" illustrate how users engage with products through questions and conditional reasoning. These are complemented by practical references to product use and effects, including "*makan*", "*minum*", "*kulit*", "*air*", "*serum*", "*booster*", and "*vitamin*". Process-related terms like "*keputusan*", "*semula*", "*restock*", "*teks*", "*bahasa*", "*perkataan*", "*bercakap*", and "*kontraksi*" reveal broader concerns about decision-making, product cycles, and communication. Emojis such as "*sparkling_heart*", "*loudly_crying_face*", and "*persevering_face*" appear but function more to adjust tone than to express strong polarity. Taken together, Neutral comments represent an evaluative phase where users seek clarification, weigh suitability, and discuss usage before forming a final opinion or making a purchase.

Table 5.3
Neutral Top Words

Top Words	Total Count
boleh	2044
mana	1788
kalau	1655
mahu	1567
tanya	1392

adakah	1321
sesuai	1278
makan	1193
minum	1147
kulit	1074
air	991
serum	938
booster	876
jerawat	854
vitamin	813
keputusan	765
semula	724
restock	699
teks	668
bahasa	641
perkataan	607
bercakap	589
kontraksi	562
sparkling_heart	517
loudly_crying_face	493
persevering_face	451
boleh	2044

Negative sentiment, as shown in Table 5.4 is characterized by highly emotional expressions, both textual and symbolic. Emojis like “loudly_crying_face”, “crying_face”, “pleading_face”, “mending_heart”, and “anxious_face_with_sweat” amplify disappointment and frustration. Supporting this, words such as “menangis”, “pusing”, “sakit”, and “bahaya” convey personal discomfort, health risks, or adverse reactions. A critical aspect of the classification process involves the model's ability to distinguish these terms within their specific syntactic context. For instance, while “sakit” is a frequent term in Negative sentiment, the GPT-4o model correctly differentiates it from phrases like “tak sakit” (not painful), which indicates a positive or neutral experience, and “sakit tak” (is it painful?), which is identified as a neutral inquiry. This ensures that the presence of the keyword does not automatically trigger a negative label but is instead interpreted through contextual negation and intent recognition. Other terms like “palsu” highlight concerns about authenticity and trust,

while “terkejut”, “cirit”, and “pemasaran” reflect negative experiences, ranging from unpleasant surprises to criticism of promotional strategies. Together, these terms show that Negative comments are often rooted in dissatisfaction, distrust, and negative personal outcomes. Compared to Positive and Neutral sentiments, which emphasize admiration or inquiry, Negative sentiment underscores the risks and frustrations that can undermine consumer confidence.

Table 5.4
Negative Top Words

Top Words	Total Count
loudly_crying_face	3798
pleading_face	768
crying_face	106
sakit	222
menangis	117
pusing	93
bahaya	87
palsu	82
terkejut	79
cirit	72
pemasaran	67
mending_heart	65
anxious_face_with_sweat	61

5.4.2.3 Sentiment Analysis for each Influencer

The stacked bar chart in Figure 5.4 shows how each of the 20 influencers’ comment sections differ in terms of Positive, Neutral, and Negative sentiment. Some influencers like Influencer 09 and Influencer 14 display a high volume of Positive feedback and relatively few Neutral or Negative comments. This pattern may result from engaging content strategies, such as viral challenges, product demonstrations, or strong personal branding that resonates with viewers. Other influencers have a more balanced sentiment profile, suggesting a mix of enthusiastic supporters, curious onlookers, and skeptical or dissatisfied users. Influencer 06, for instance, exhibits a greater proportion of Negative comments, which could be attributed to unmet expectations or a recent controversy. Influencer 13, on the other hand, has minimal

engagement overall, indicating limited audience interaction or lower visibility. Understanding influencer-specific sentiment distribution helps identify best practices for content creation and brand alignment. Highly positive influencers can leverage their existing goodwill, while those facing neutral or negative feedback may need to address user concerns more directly such as potentially refining their messaging, clarifying product information, or engaging in transparent communication to rebuild trust.

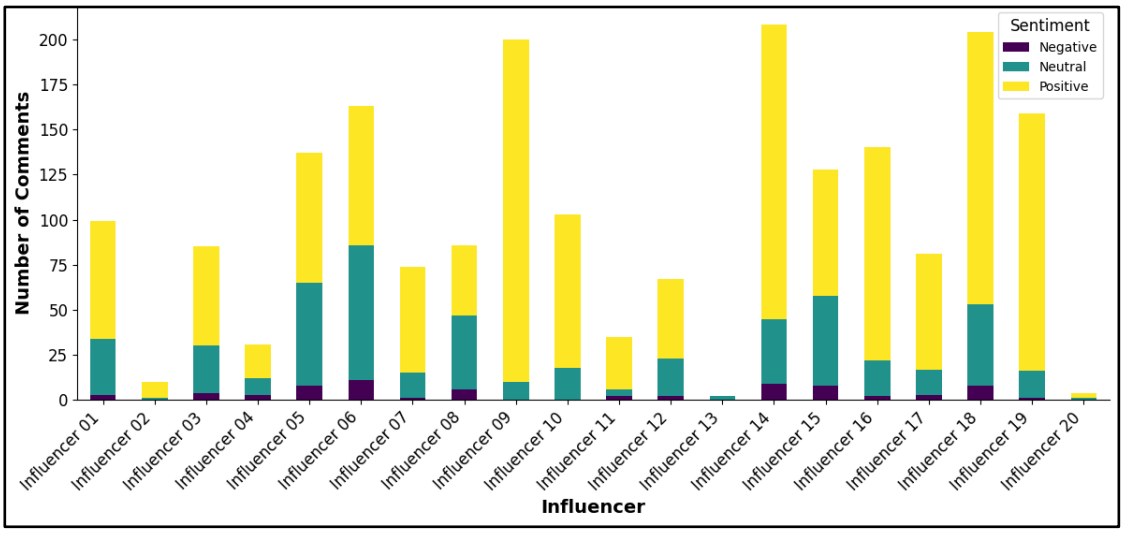


Figure 5.3 Sentiment Analysis By Influencer

5.4.2.4 Influencer Sentiment and Revenue

Table 5.2 presents a detailed breakdown of Negative, Neutral, and Positive comment counts for each influencer, alongside their corresponding revenue. While it might be assumed that high Positive sentiment correlates with higher earnings, the data reveals a more nuanced relationship. For example, Influencer 02 demonstrates a small number of comments yet reports the highest revenue (RM1,540,000), suggesting that brand collaborations or exclusive sponsorships can yield substantial financial returns despite relatively low audience engagement. Conversely, Influencer 09 garners an impressive number of Positive comments yet earns considerably less (RM322,980). This discrepancy indicates that sentiment alone may not drive revenue; additional factors such as live selling tactics, brand partnerships, or direct product endorsements can significantly influence an influencer's income. Neutral comments play a pivotal role, as they often signal prospective buyers seeking product details or clarifications—an overlooked but valuable engagement opportunity. A high volume of Positive

sentiment is not a guaranteed predictor of higher revenue. Instead, influencers and marketers should focus on converting Neutral inquiries into actual purchases, maintaining transparency to address any Negative feedback, and exploring monetization strategies that align with audience trust and content engagement.

Table 5.5
Influencer's Distribution Of Sentiment And Their Corresponding Revenue

Influencer	Negative (Count)	Neutral (Count)	Positive (Count)	Revenue (RM)
Influencer01	3	31	65	302240
Influencer02	0	1	9	1540000
Influencer03	4	26	55	429780
Influencer04	3	9	19	330480
Influencer05	8	57	72	305310
Influencer06	11	75	77	311150
Influencer07	1	14	59	545570
Influencer08	6	41	39	451530
Influencer09	0	10	190	322980
Influencer10	0	18	85	346310
Influencer11	2	4	29	361520
Influencer12	2	21	44	348890
Influencer13	0	2	0	328490
Influencer14	9	36	163	446480
Influencer15	8	50	70	485200
Influencer16	2	20	118	418690
Influencer17	3	14	64	548190
Influencer18	8	45	151	577870
Influencer19	1	15	143	593260
Influencer20	0	1	3	304490

5.4.3 Actionable Insights

The sentiment analysis indicates that Positive sentiment dominates, although Neutral comments also play a critical role in audience engagement, and Negative sentiment is minimal. Purchase-related words in Positive comments reflect consumer excitement for products and influencers, highlighting that enthusiasm and admiration drive much of the interaction. At the same time, Neutral comments often involve

product inquiries, demonstrating that users engage not only on an emotional level but also for informational purposes. This pattern underscores the value of responding to factual or question-based comments through interactive Q&A sessions, influencer-led product demonstrations, and active comment engagement, all of which can help convert interest into sales. Although Negative sentiment is relatively low, it still reveals important issues like unmet expectations or product concerns, indicating that brands and influencers should remain vigilant in addressing dissatisfaction promptly to sustain credibility.

Emojis emerge as vital sentiment indicators, underscoring the importance of non-verbal expressions in TikTok's informal communication style. Positive sentiment is frequently tied to joyful emojis and explicit buying interest, while Neutral sentiment leans on factual words that drive decision-making processes, and Negative sentiment often involves dissatisfaction-related terms and crying-face emojis. This highlights the need for sentiment models that integrate multimodal understanding, including emoji interpretation, to avoid misreading user intentions. Across various influencers, the share of Neutral comments fluctuates, particularly among those offering educational content or product reviews. Revenue data further illustrates that sentiment alone does not determine earnings: some influencers with high Positive sentiment earn less than expected, whereas others with fewer comments achieve substantial revenue through strategic partnerships, live selling, or brand collaborations. The capacity to convert Neutral sentiment into purchases stands out as a pivotal factor in financial success, emphasizing the importance of fostering trust, addressing user inquiries, and aligning content with consumer expectations.

5.5 Conclusion

In conclusion, the GPT-based sentiment analysis in this chapter highlights a predominantly Positive audience in TikTok's Beauty and Personal Care domain, with Neutral comments emerging as a critical yet sometimes underappreciated driver of user engagement and potential conversions. Although the volume of Negative feedback remains low, it underscores the importance of addressing product or content concerns to sustain credibility. Moreover, the correlation between sentiment and influencer revenue is not strictly linear, emphasizing that monetization strategies, brand

collaborations, and responsiveness to consumer inquiries all play pivotal roles in financial outcomes.

By integrating these insights, influencers and brands can tailor their TikTok content strategies to foster user trust, encourage more interactive engagements, and convert neutral or inquisitive viewers into active customers. Critically, this chapter addresses the core problem statement regarding the difficulty of extracting business intelligence from linguistically complex social media data. By demonstrating that GPT-4o can successfully navigate Malaysian slang, code-switching, and nuanced syntax such as distinguishing between negations ("tak sakit") and inquiries ("sakit tak"), this study contributes a validated, replicable workflow for sentiment analysis in the Southeast Asian digital context.

While this chapter has successfully identified how consumers feel about the content (the sentiment polarity), it does not yet explain what specific aspects they are discussing. Understanding that a comment is "Positive" is valuable, but it is insufficient for strategic decision-making without knowing the underlying context whether the positivity is directed at the product's efficacy, its packaging, or the influencer's presentation style. To bridge this gap and address the remaining research objectives, Chapter 6 will advance the analysis from sentiment detection to Topic Modeling. By employing the BERTopic algorithm, the next chapter will uncover the latent themes and discussion patterns within the comments, thus providing the necessary context to fully interpret the drivers of the sentiments identified in this chapter.

CHAPTER 6
TOPIC MODELING

6.1 Introduction

This research applies a topic modeling technique to TikTok user comments in the Beauty and Personal Care category. The goal is to identify common themes and patterns in how users talk about products, experiences, and preferences. By grouping similar comments, we can detect recurring topics that matter to users in which can support content strategy, customer feedback analysis, and marketing decisions. In this research, a specific sequence of steps to conduct the modelling is followed. The process begins with data preprocessing, and machine learning approach called BERTopic to generate and interpret the topics. Figure 6.1 shows the full process, starting from raw data to final topic generation. Each step plays a role in cleaning, transforming, and organizing the comment data so that meaningful patterns can be extracted.

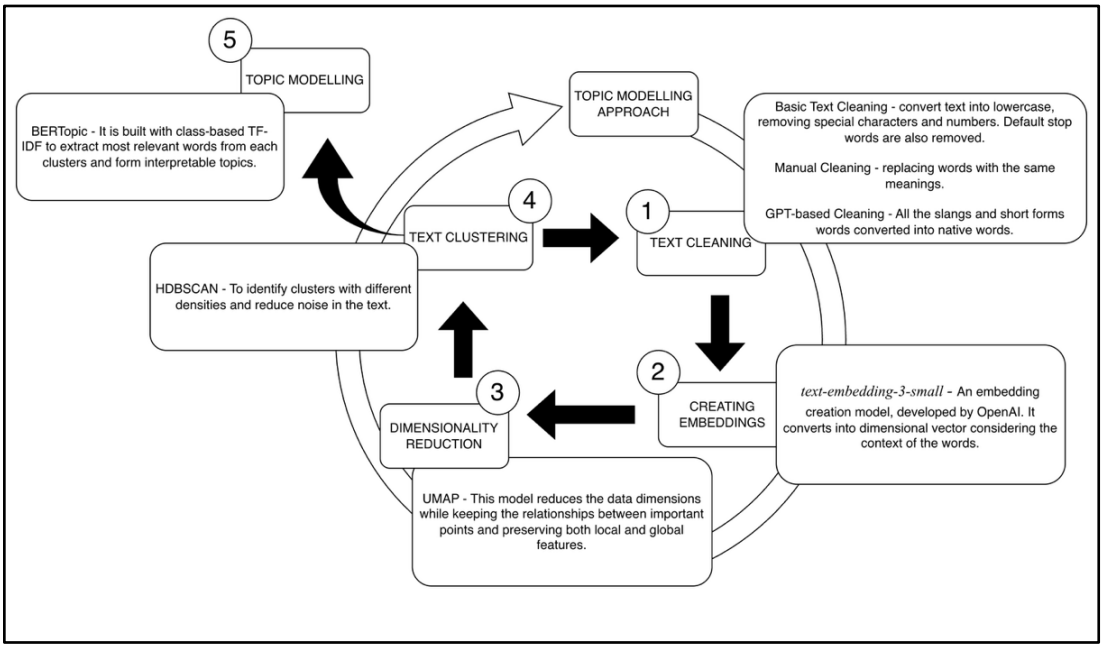


Figure 6.1 Full Topic Modeling Process

6.2 Text cleaning

The first step in the topic modelling process involves text cleaning, which is essential to prepare the raw data for further analysis. Even though the comment text is processed in the data preparation phase, a specific cleaning process needs to be conducted when performing topic modelling. The text cleaning process is carried out through several specific procedures tailored to handle the TikTok comments complexity. First, all the text is converted to lowercase to ensure consistency in the dataset, as this prevents the word like "*Cantik*" and "*cantik*" from being treated differently. This help simplify the vocabulary and improves the efficiency of subsequent analysis (Hickman et al., 2022). Next, non-alphabetic characters which encompass numbers and special symbols are removed to reduce noise. Since these characters does not add much value to the semantic meaning of the text. In this research, it is important to exclude the special characters and numbers to keep the analysis focused on the textual content (Banks et al., 2018). Even though sometimes punctuation can indicate emotional tone, but such features are not essential for this research.

Additionally, stop words, which are common words like "*yang*" and "*dan*" that does not carry significant meaning in the context of topic modeling, are also removed. This step reduces the number of unimportant terms, enabling the analysis to focus on the more meaningful words that are central to the themes in the dataset (Vijayarani & Research Scholar, 2015). Short words that are less than 3 characters are also excluded to further streamline the text and avoid trivial terms cluttering the dataset. Some manual adjustments are also made to the dataset. Words that are overly common or specific to the domain but do not contribute significant meaning, such as generic terms or redundant phrases, are also removed. Furthermore, synonymous words are consolidated to prevent redundancy, ensuring that the dataset is more concise and relevant for analysis.

6.3 Modeling Approach

6.3.1 Embeddings

After cleaning the TikTok comments, the next step is to convert the text into a numerical format that can be used in machine learning models. This process is called embedding, and it is a common approach in natural language processing (NLP) to

represent text data as vectors. In this research, the embedding model used is "*text-embedding-3-small*" by OpenAI. This model transforms each comment into a 1,536-dimensional vector. Each vector captures the semantic meaning of the comment which reflects what the comment is about and not just what words are in it. Compared to traditional methods like Bag of Words or TF-IDF, embeddings are more effective because they consider the context of the words (Grootendorst, 2022; Shanbhag et al., 2025). For example, the word "*bau*" might appear in different comments, but its meaning could change depending on surrounding words. Embedding models are trained to capture these subtle differences in meaning.

This embedding step is especially important for short and informal content like TikTok comments. Because the comments are often brief, slang-heavy, or emoji-based makes traditional methods struggle to find meaningful patterns (Amur et al., 2023). By using a transformer-based embedding model, this research leverages a more advanced method that performs well even with limited text. Embedding is a foundational step in many Natural Language Processing (NLP) applications including chatbots, semantic search engines, recommendation systems, and social media analytics (Zhao et al., 2023). By turning TikTok comments into vectors, we prepare the data for further processing such as clustering and topic extraction in the next steps.

6.3.2 Dimensionality Reduction

In this research, dimensionality reduction phase plays a crucial role to manage high dimensional data, especially given the complexity of the TikTok comments dataset. Handling a high-dimensional data could lead to the "*curse of dimensionality*", where analysis process becomes inefficient and less accurate as the number of features increases (Mekala et al., 2019). To manage this, the Uniform Manifold Approximation and Projection (UMAP) technique was used. UMAP is a dimensionality reduction method designed to preserve the underlying structure of the data. It reduces the data into fewer dimensions while keeping the important relationships between points (Ghojogh et al., 2021). UMAP was applied with "*n_components = 5*" and "*metric = 'cosine'*". These settings help retain both local and global patterns in the data, which is important when working with informal and varied content like TikTok comments. UMAP also handles noisy data better than traditional methods like PCA, making it a good fit for user-generated content.

Dimensionality reduction is a common step in NLP workflows, especially in machine learning and data visualization tasks. It helps simplify large and complex data so that downstream processes such as clustering can work more effectively. By using UMAP, the model produces cleaner input for the next stage of topic modeling.

6.3.3 Clustering

Nex step in topic modelling process involves clustering the reduced embeddings to group comments based on their semantic similarities. This clustering step plays a vital role in organizing the data into coherent groups that share common underlying themes. For this step, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is deployed, as it builds upon the strengths of DBSCAN and handles noise and outliers more effectively (Scoccola & Rolle, 2023b). HDBSCAN clusters the embeddings by calculating the core distance for each data point, constructing a minimum spanning tree from a weighted distance graph, and then clusters the data based on density (Malzer & Baum, 2019).

One of the key advantages of utilizing HDBSCAN over other traditional clustering algorithms is its ability to automatically determine the optimal number of clusters and identify outliers which is data points that did not fit into any cluster. This is important when working with TikTok data, where many comments may be short, duplicated, or off-topic. By using HDBSCAN, the model can focus on semantically meaningful patterns while filtering out noise. Clustering is a standard unsupervised learning approach used in applications like customer segmentation, anomaly detection, and content organization. Here, it plays a critical role in structuring unstructured text data, enabling the next phase, which is topic modeling, to work on high-quality and well-separated groups of related comments to produce interpretable topics.

6.4 Topic Generation

Once the embeddings were clustered, the next step was to generate and interpret the topics. This was done using BERTopic, a topic modeling technique that works well with short and informal text, such as social media comments. BERTopic combines clustering results with a class-based TF-IDF approach to extract the most relevant words from each cluster and form interpretable topics. In this research, BERTopic was applied

to the output of the HDBSCAN clusters. Each cluster was treated as a topic, and the most representative words in each topic were selected based on their importance across the dataset. To improve clarity, the model was configured to use a trigram-based tokenizer, which helps detect common phrases in user comments (e.g., “bau wangi sangat”). From the 34,597 TikTok comments analyzed, two valid and coherent topics were identified after filtering out noise and irrelevant clusters. Table 6.1 below shows the two topics reflect the main areas of interest in the Beauty and Personal Care category.

Table 6.1
Topics Generated

Topic	Themes and Description	Top Words
0	Product Usage and Experience: This topic includes comments related to how users use the product, what they feel after using it, and how effective they think it is.	“pakai”, “ada”, “yang”, “rambut”, “dengan”, “kalau”, “untuk”, “cantik”
1	Product Features and Attributes: This topic focuses on specific aspects of the product such as smell, packaging, and visual appeal.	“bau”, “wangi”, “botol”, “minyak wangi”, “pesanan”

The following visual illustrates the top words associated with each topic, based on their term frequency–inverse document frequency (TF-IDF) scores.

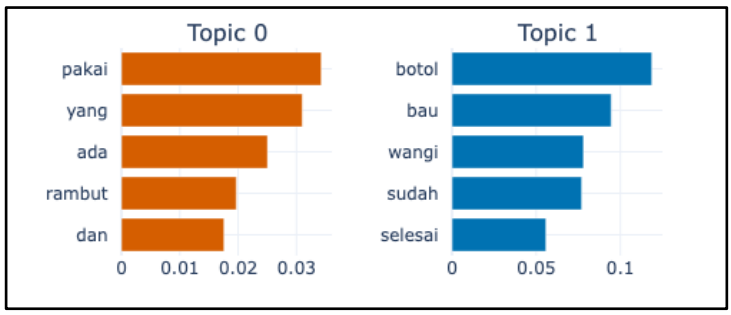


Figure 6.2 Top Words In Each Topic By TF-IDF Score

The figure 6.2 above shows the top five keywords for each of the two generated topics. On the left, Topic 0 includes words such as *pakai*, *yang*, *ada*, *rambut*, and *dan*, which are tied to how users describe applying the product or observing results. This suggests the topic captures user routines, effectiveness, and outcomes. On the right, Topic 1 includes *botol*, *bau*, *wangi*, *sudah*, and *selesai*, indicating that users in

this group are commenting on the product's design, fragrance, or packaging experience. The distribution of keywords shows a clear distinction in focus between the two topics, validating that the model has successfully segmented different types of user discussions.

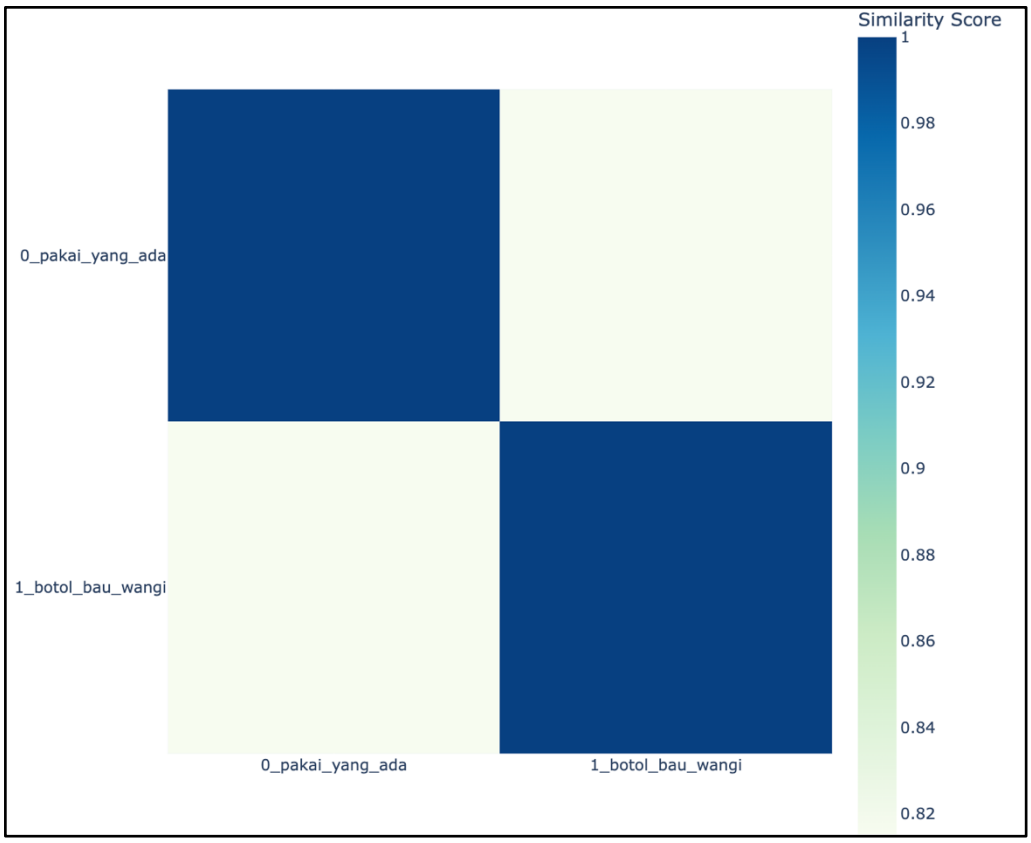


Figure 6.3 Similiarity Matrix

The similarity matrix in figure 6.3 above displays the cosine similarity scores between the two topics. As expected, the diagonal values are 1, indicating perfect self-similarity. The off-diagonal score approximately 0.82 represents the semantic similarity between Topic 0 and Topic 1. This relatively high score shows that while the topics are distinct, they are still related under the broader context of product engagement. This matrix is useful for validating whether clustering and topic modeling results have successfully differentiated content areas. This helps in building structured data representations for further processing in recommender systems, feedback analysis pipelines, or trend detection platforms.

6.5 Topic Coherence

In this research, a method to evaluate how well the topics generated by BERTopic is employed to decide whether the topic represent meaningful and coherent ideas. This method is known as topic coherence, which measures how closely the words within a topic are related to one another, forming a clear and understandable theme. The higher the topic coherence score, the more logically connected and easier to interpret the topic becomes (Bianchi et al., 2020). Several metrics are applied to assess this coherence. One of the key metrics is c_v, which combines the normalized pointwise mutual information (NPMI) and cosine similarity to evaluate how often words within a topic appear together. Another commonly used metric is u_mass, which measures word co-occurrence based on document frequency. Additionally, c_uci examines the co-occurrence of word pairs, and c_npmi focuses on the normalized pointwise mutual information, which helps gauge the overall coherence of the topics (Campagnolo et al., 2022).

These metrics are crucial to ensure that the topics generated from the data are both relevant and easily interpretable. However, the effectiveness of these metrics can vary depending on factors such as the number of topics, the size of the dataset, the vocabulary used, and the text preprocessing techniques applied. By using these methods, the research ensures that the topics are meaningful and provide useful insights. The scores obtained in this research are shown below.

Table 6.2
Coherence Score

Metric	Score	Interpretation
c_v	0.620	High coherence, especially for short social media text.
u_mass	-1.009	Acceptable in noisy, informal datasets.
c_uci	0.918	Strong word-pair association.
c_npmi	0.179	Moderate semantic consistency.

The c_v score of 0.620 suggests a strong level of coherence across the top terms in each topic, making the results interpretable and useful for further analysis. While u_mass is slightly negative, this is typical when working with short-form user-generated content, where full sentence context is limited. The c_uci score (0.918) supports the presence of meaningful word groupings, and the c_npmi score (0.179)

indicates moderate topic consistency across the corpus. These results confirm that the topic modeling pipeline, which combines transformer embeddings, UMAP, HDBSCAN, and BERTopic, produced logically coherent and semantically valid topics which suitable for use in downstream IT systems such as trend detection tools, social media intelligence platforms, or text-based recommender systems.

6.6 Actionable Insights

The results from the topic modeling process reveal two dominant and interpretable themes in TikTok user comments: (1) product usage and experience, and (2) product features and attributes. These topics provide practical insights that can be directly applied to Information Technology solutions, particularly in systems focused on content analysis, user feedback interpretation, and data-driven strategy support. First, the comments categorized under Topic 0 focus on how users describe their interaction with a product including how often they use it, the results they observe, and their level of satisfaction. From an IT perspective, this type of information is valuable for systems that support automated sentiment tracking or real-time feedback dashboards. For instance, e-commerce platforms or marketing teams can integrate this data into dashboards that show aggregated user experience metrics for specific product lines.

Second, Topic 1 highlights the importance of product-related features such as scent, packaging, and presentation. This reflects how aesthetic and sensory factors influence user perception, which is especially relevant in the beauty and personal care domain. These insights can be integrated into social listening tools or product design feedback loops, where technical teams can monitor user focus on non-functional aspects that affect product appeal. In practice, these themes could support various Information Technology driven applications, such as:

- Recommendation engines that prioritize products discussed positively in both usage and sensory terms.
- Customer service automation, where common concerns or praise points from each topic can be converted into FAQ entries or chatbot responses.
- Influencer content analytics, identifying which video topics are aligning with the audience's discussion trends.
- Marketing intelligence systems that segment user-generated content into

emotional (experience-driven) and rational (feature-driven) categories.

The clear separation between these two topics also allows future systems to build modular analytics pipelines, for example, tracking changes in product usage sentiment over time while separately analyzing attention to packaging or delivery. In this way, topic modeling goes beyond raw analysis and becomes a tool for driving structured, technology-supported decisions

6.7 Conclusion

In conclusion, this chapter presented the full topic modeling process applied to TikTok comments in the Beauty and Personal Care category, using a transformer-based NLP pipeline. Starting from text cleaning, the research applied GPT-based embeddings, dimensionality reduction with UMAP, clustering with HDBSCAN, and topic generation using BERTopic. The empirical contribution of this modeling process is the identification of two distinct and high-quality thematic pillars which are Product Usage and Experience (Topic 0) and Product Features and Attributes (Topic 1). These findings provide a structured understanding of consumer priorities, revealing how users distinguish between functional outcomes and sensory attributes like scent or packaging. The reliability of these contributions is validated by strong coherence metrics, specifically a c_v score of 0.620 and a c_uci score of 0.918, which confirms that the generated topics are both logically consistent and semantically valid. Ultimately, this chapter contributes a practical framework for converting large-scale and unstructured user-generated content into structured data. This transition is essential for driving decision-support in Information Technology systems, such as real-time feedback dashboards, automated customer service responses, and modular marketing intelligence pipelines. By bridging the gap between raw social media data and actionable IT insights, this research addresses a core challenge in the digital platform economy.

CHAPTER 7
DASHBOARD

7.1 Introduction

Dashboards play a crucial role in Information Technology by serving as interactive interfaces that transform complex data outputs into accessible insights. In this research, the dashboard visualizes results from sentiment analysis, topic modeling, and influencer engagement analysis based on TikTok user-generated content. The dashboard acts as the final component of the data pipeline, integrating processed outputs into a unified platform for decision support. Built using visualization tools, it supports key Information Technology principles such as system integration, human-computer interaction (HCI), and real-time data exploration. In addition to its design and implementation, this chapter also presents the results of usability testing, which evaluates the dashboard's clarity, navigation, and overall user experience. Together, these sections highlight the dashboard's role in enabling data-driven strategies for social media analytics.

7.2 System Architecture and Tools

The dashboard developed in this research is the final component of a modular data pipeline that integrates multiple data analytics processes into a unified interface for decision support. Its architecture reflects core principles in Information Technology, particularly in data engineering, system integration, and interactive visualisation. The system is designed to transform unstructured TikTok user-generated content (UGC) into structured, meaningful insights through natural language processing (NLP), sentiment analysis, and topic modelling. At a high level, the system follows a multi-phase architecture. Data acquisition was handled using Apify, which extracted TikTok metadata and comments from selected influencers. The raw data was then pre-processed using python, including GPT-based text normalization and cleaning, emoji-to-text conversion, and multilingual handling (Kheiri & Karimi, 2023; Yenduri et al., 2024). Sentiment classification was performed using the GPT-4o API, while topic modelling

leveraged BERTopic, which combined transformer0based embeddings with UMAP and HDBSCAN for dimensionality reduction and clustering.

Processed outputs were structured into tabular formats (CSV/Parquet), which served as the data source for the dashboard layer. The dashboard itself was built using Streamlit, selected for its robust data integration features, real-time filtering capabilities, and support for interactive visual components (Puchakayala et al., 2025). This platform enabled dynamic linking of visual modules like sentiment charts, topic distributions panels, and influencer engagement metrics, and allowed end users to filter and drill down across dimensions such as topic label and sentiment class (Bach et al., 2022; Ruoff et al., 2023). The system architecture illustrated in Figure 7.1, showing a data flow from raw TikTok UGC to the final interactive dashboard. The modular nature of this pipeline supports reusability, maintainability, and extensibility which are essentials for modern IT systems (Macías & Borges, 2024). Furthermore, this approach aligns with (Bach et al., 2022)'s dashboard design principles by enabling layout control, interactivity, and visual composition using both content and composition patterns. Therefore, the tools and architecture in this research emphasizes scalability, user interaction, and data interpretability to enable users to derive real-time, actionable insights from complex social media data streams (Macías & Borges, 2024; Ruoff et al., 2023).

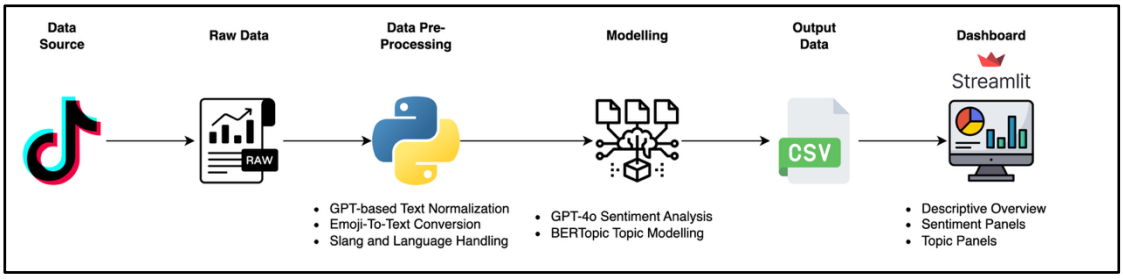


Figure 7.1 Data Flow

7.3 Dashboard Design Objectives

The dashboard developed in this research is designed to serve as an interactive decision support tool that transforms complex analytics outputs into actionable business insights. Functionally, it enables users to explore sentiment distributions, interpret topic modelling results and evaluate influencer performance using visual summaries derived from unstructured TikTok user-generated content. The design focuses on supporting

users such as brand managers, influencers and marketers by allowing them into monitoring trends in consumer feedback, assess content engagement, and identify emergent themes related to product perception. These objectives align with the core functions of modern dashboards in decision-critical environments, where monitoring, comparison, and interpretation must be both rapid and reliable (- et al., 2022; Frazao et al., 2021; Puchakayala et al., 2025).

From an Information Technology (IT) perspective, the dashboard embodies key principles of modular system design. It is built on a decoupled architecture that separates data collection, natural language processing, and visualisation into distinct yet interoperable layers. This modularity supports reusability and scalability which are critical traits in IT systems that must evolve with new data sources, models, or deployment platforms (Almadani et al., 2025; Joseph et al., 2020). For example, the underlying sentiment analysis and topic modelling logic can be updated independently of the visual interface. Furthermore, the dashboard implements data abstraction by summarizing complex transformer-based NLP outputs like GPT-based sentiment classes and BERTopic clusters into simplified visuals which could make these results interpretable for non-technical users. These features reflect the design trade-offs discussed by (Bach et al., 2022), where screen space, abstraction, and interactivity must be carefully balanced to maximize usability and insight communication in data-intensive dashboards.

In line with Human-Computer Interactions (HCI) best practices, the dashboard emphasizes usability, clarity and low cognitive load. Visual elements such as bar charts, sentiment-coloured labels, and topic word clouds are arranged using a clean layout to guide user attention and facilitate intuitive navigation. Filtering capabilities allow users to drill down by sentiment class or topic to enables a dynamic insights retrieval without overwhelming the interface (Chowdhury et al., 2021; Islam et al., 2024). This approach reflects the need for user-centred design in decision support systems, particularly those intend for business contexts involving diverse user roles and variable data literacy. Furthermore, recent advancements in conversational dashboards (Ruoff et al., 2023) suggest promising extensions to this work like integrating chatbot-like querying mechanisms to make complex analyses even more accessible through natural language interfaces.

Overall, the dashboard is not only a visualisation tool but also a functional representation of IT contributions to data interpretation, system integration, and

business intelligence. It demonstrates how interactive systems can operationalise artificial intelligence (AI) model outputs and deliver then through interfaces that are technically robust, user-centric, and strategically aligned with real-world decision-making needs.

7.4 Dashboard Components and Visual Modules

The dashboard implemented in this research consists of several interactive modules, each representing distinct dimensions of insights derived from the TikTok dataset. These modules are structured into logical sections, enabling layered exploration from high-level metrics to detail text analytics. This section details the design of each visual component, the underlying data logic, and how they contribute to business intelligence interpretation through an interactive Streamlit-based interface.

7.4.1 Revenue Summary and Influencer Profiling

The dashboard begins with a revenue summary section that aggregates and visualises influencer-level financial performance. Key metrics such as total influencers, average revenue, and average followers are displayed using visually centred cards styled with HTML and CSS embedded in Streamlit (as in Figure 7.2). These cards enable immediate executive-level insights into the influencer landscape.

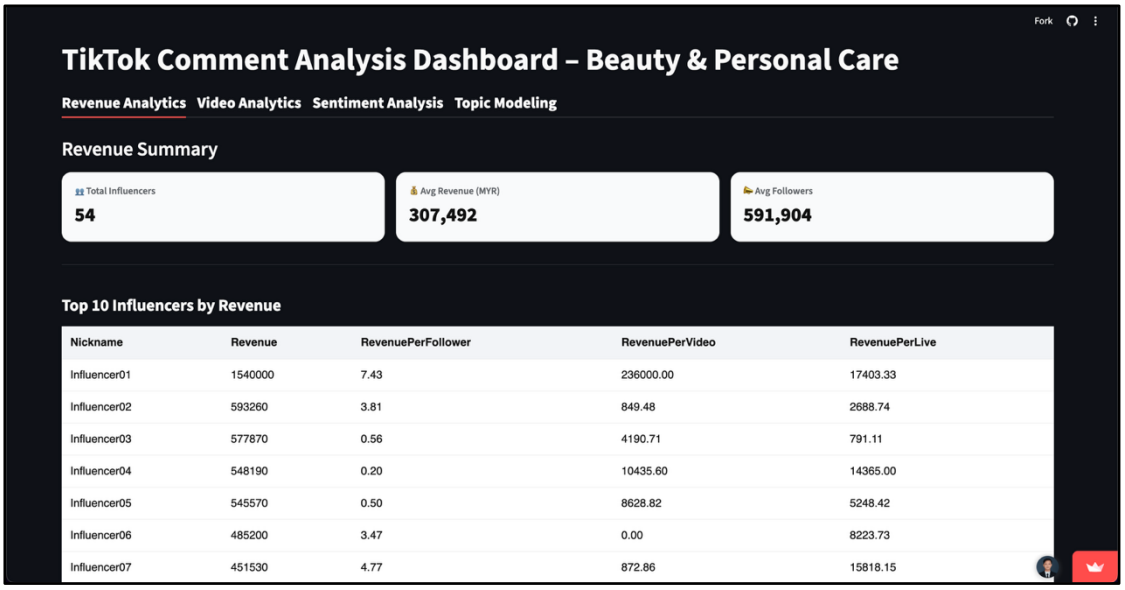


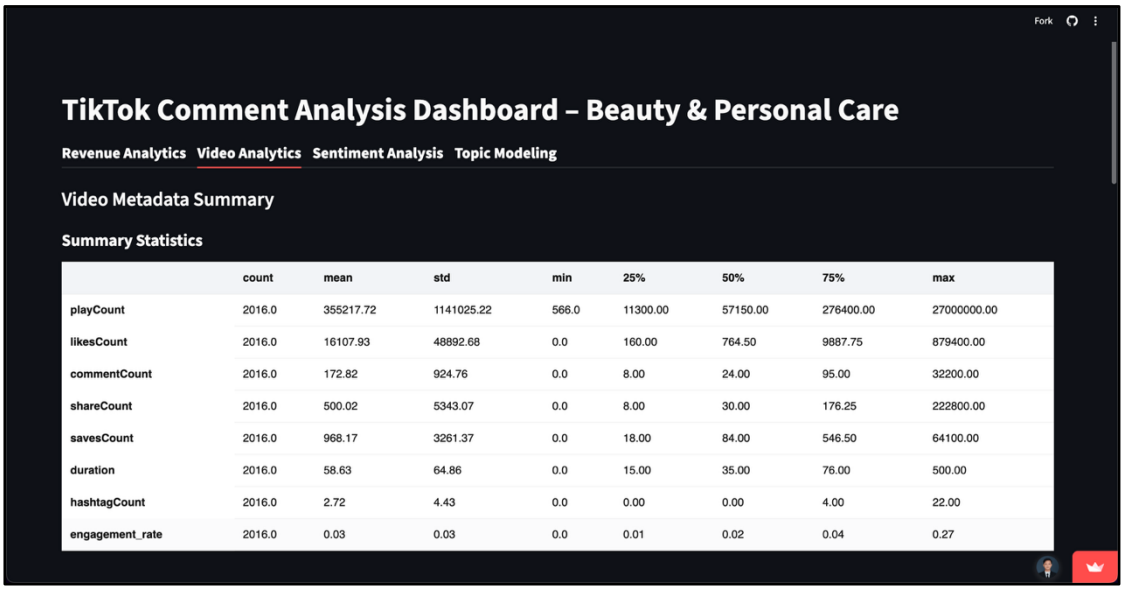
Figure 7.2 Revenue Summary

Below the cards, a table ranks the top 10 influencers by revenue, with derived metrics like Revenue per Follower, Revenue per Video, and Revenuer per Live calculated and formatted for comparison (as in Figure 7.2). This module aids business users in identifying high-performing influencers and benchmarking their monetization efficiency. A correlation heatmap follows in this section (Figure 7.3), illustrating relationships between multiple variables, such as video count, followers and various revenue components. This component implemented using "*seaborn.heatmap*" to support exploratory data analysis and hypothesis generation.

Figure 7.3 Revenue Correlation Matrix

7.4.2 Metadata Summary and Engagement Patterns

The metadata summary module (as in Figure 7.4) presents descriptive statistics of key engagement metrics such as views, likes, comments, and share counts. These statistics are rendered in styled HTML tables and help users understand the overall scale and variance within the TikTok dataset.



	count	mean	std	min	25%	50%	75%	max
playCount	2016.0	355217.72	1141025.22	566.0	11300.00	57150.00	276400.00	27000000.00
likesCount	2016.0	16107.93	48892.68	0.0	160.00	764.50	9887.75	879400.00
commentCount	2016.0	172.82	924.76	0.0	8.00	24.00	95.00	32200.00
shareCount	2016.0	500.02	5343.07	0.0	8.00	30.00	176.25	222800.00
savesCount	2016.0	968.17	3261.37	0.0	18.00	84.00	546.50	64100.00
duration	2016.0	58.63	64.86	0.0	15.00	35.00	76.00	500.00
hashtagCount	2016.0	2.72	4.43	0.0	0.00	0.00	4.00	22.00
engagement_rate	2016.0	0.03	0.03	0.0	0.01	0.02	0.04	0.27

Figure 7.4 Metadata Summary

A second correlation matric shows how these engagement variables interact by providing visual insights into how features like hashtag count and duration affect engagement rate as shown in Figure 7.5. Complementing this, a bar chart of average engagement rate by hashtag count is implemented with *Plotly Express* to showcase that higher hashtag usage does not linearly increase engagement which is a valuable insight for marketing strategy.

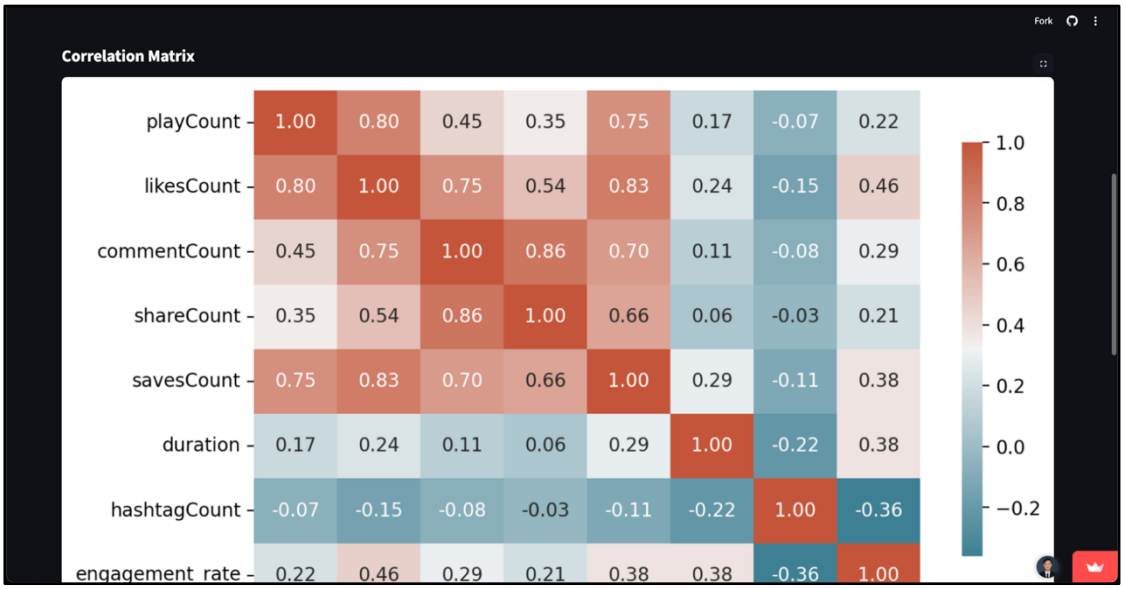


Figure 7.5 Metadata Correlation Matrix

7.4.3 Sentiment Overview

The sentiment overview module (as in Figure 7.6) displays the distribution and count of sentiments (Positive, Neutral, Mixed, and Negative) across TikTok comments. Users can interactively highlight a specific sentiment class to emphasize or de-emphasize it visually, which dynamically adjust both the pie chart and bar chart outputs using "*matplotlib*". This functionality provides real-time sentiment feedback and allows users to understand the emotional tone of consumer discourse. Contextual summaries are also included to clarify data sample size and label definitions.

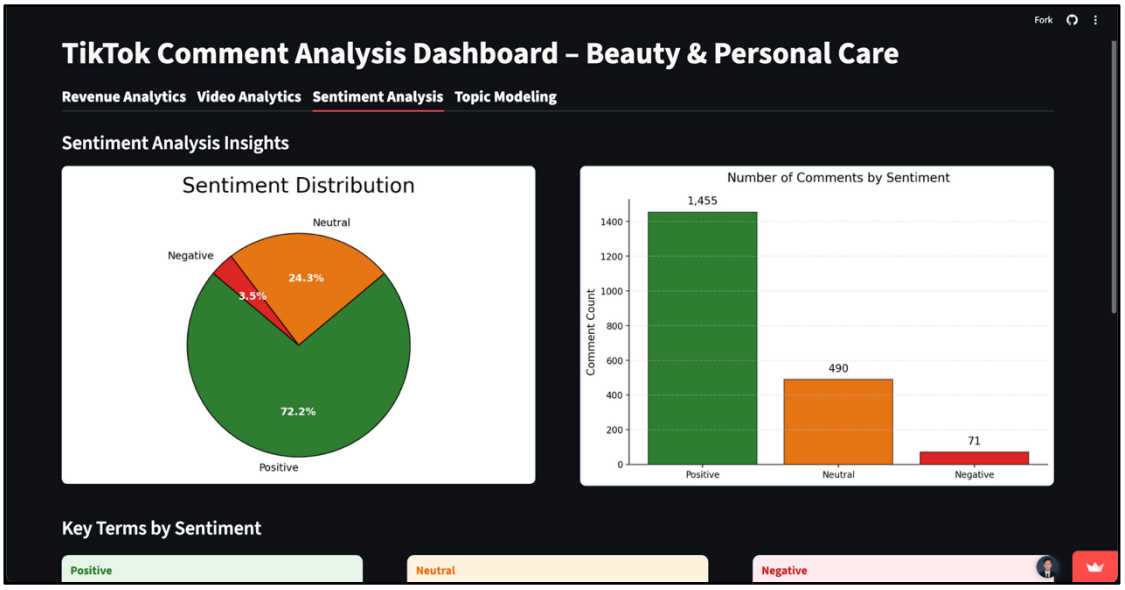


Figure 7.6 Sentiment Overview

7.4.4 Topic Modelling Overview

The final major component is the topic modelling insights section. This module displays a bar chart of topic frequencies, showing how many comments belong to each topic. For each topic, a paired display of word clouds (generated from representative comments), top keywords (extracted from topic representations), and sample comments (drawn directly from the BERTopic pipeline outputs. There are two dominant themes identified from the dashboard which are general product usage and effectiveness, and packaging and scent preferences. Each of these is visually separated and labelled by topic index and name (refer Figure 7.7 and 7.8). These modules transform unstructured

comment text into digestible clusters for consumer insight, enhancing interpretability for non-technical users through natural language summarisation and visual anchoring.

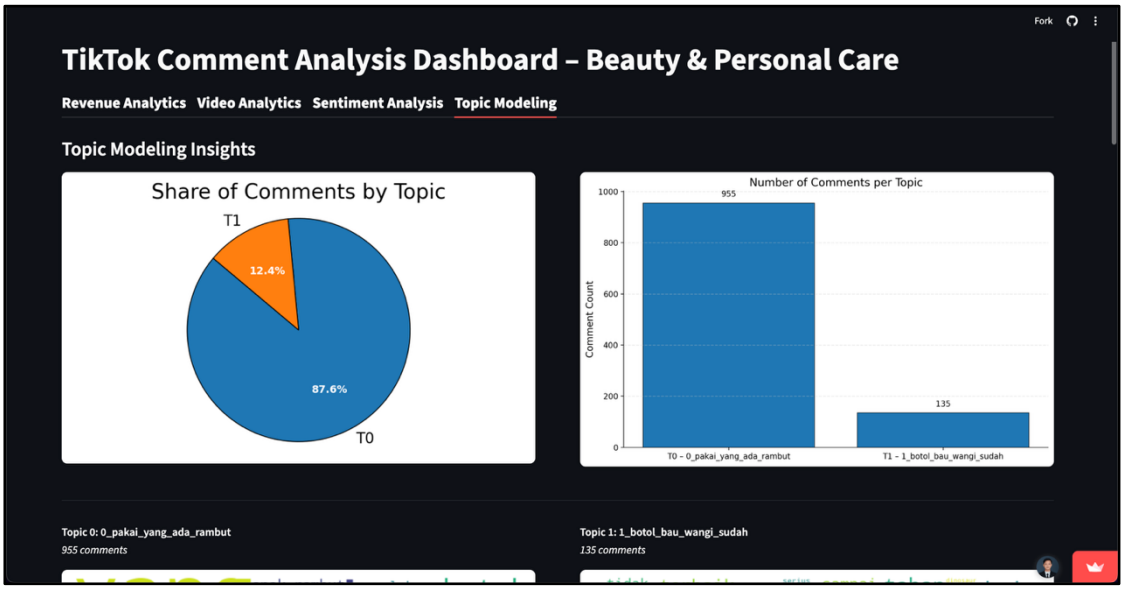


Figure 7.7 Overview of Topic

Figure 7.8 Context for each Topics

7.5 User Experience

While interactivity features are limited in the current Streamlit dashboard, the overall user experience remains a key design priority. This section discusses two

important aspects of user experience, visual clarity and cognitive load and accessibility and portability of the system.

7.5.1 Visual Clarity and Cognitive Load Management

Effective dashboards are not only measured by the volume of data presented but also by how intuitively the information can be interpreted. The ensure clarity, the dashboard adopts a modular layout strategy, organizing insights into well-defined sections of Revenue Summary, Metadata Overview, Sentiment Analysis and Topic Modelling. Each module is separated using visual elements such as headers, whitespace, and horizontal rule, which collectively reduce cognitive burden on the viewer (Ke et al., 2023). The use of column-based layout allows related visuals to be presented side by side. For example, the sentiment pie chart and corresponding bar chart are aligned horizontally to facilitate comparative reading. This dual-view technique supports mental integration of proportions and absolute counts, minimizing the need for users to switch between tabs or visual contexts (Ryu et al., 2003; Shaikh et al., 2022).

Colour semantics were carefully applied to match intuitive expectations like green for positive, red for negative, and orange for neutral, aiding for quick interpretation without the need for legend memorization. Bar plots and pie charts are also annotated directly to reduce reliance on external labels, a known technique to lower extraneous cognitive load in data visualisations (Donohoe & Costello, 2020; Hadjimichael et al., 2024). These techniques collectively support the principle of cognitive load minimization in dashboard design, allowing users to focus on insight extraction rather than interpretation mechanics.

7.5.2 User Accessibility and Portability

To ensure accessibility, the dashboard was implemented using Streamlit, a lightweight, python-based web application framework. Streamlit applications are platform-independent, allowing users to access the dashboard via any modern web browser without installation or technical configuration. This aligns with inclusive design principles and supports a wide range of users across devices. The dashboard is also designed with performance and portability in mind. All computational processes

are pre-processed offline, allowing dashboard to render insights quickly with minimal lag. The use of optimized libraries ensures graphical elements are rendered efficiently.

Although mobile optimization is partial, the dashboard remains usable on smaller screens with scrollable layouts adjustments. From a deployment perspective, the system can be hosted via various online platforms such as Stream Cloud or GitHub. In summary, despite limited interactivity features, the dashboard demonstrates effective user-centred design, balancing clarity, accessibility, and usability. This ensures that the insights derived from AI-based sentiment and topic models are presented in a form that is actionable, interpretable, and ready for decision-making, especially for non-technical users.

7.6 Usability Testing

7.6.1 Objectives and Methods

Usability testing is a well-established method for evaluating whether an interface can be effectively used by its intended audience (Almasi et al., 2023; Nielsen, 1994). For dashboards, usability is especially important because the system must communicate complex information quickly and clearly without requiring extensive training or technical guidance (Almasi et al., 2023; Richter Lagha et al., 2020). In this research, usability testing was conducted to determine whether the TikTok Analysis Dashboard provided an intuitive user experience, supported efficient navigation, and presented results in a way that users could easily interpret. This approach follows common practices in human–computer interaction, where standardized questionnaires such as the System Usability Scale (SUS) and qualitative feedback are used to evaluate interactive systems (Brooke, 2013; Issa & Isaias, 2022).

The primary objective of the test was to assess ease of use, navigation, readability of charts, clarity of insights, and overall satisfaction with the dashboard (Dolatabadi et al., 2024). Participants were first given access to the live Streamlit dashboard and asked to explore its main modules, including revenue analytics, video metadata, sentiment analysis, and topic modeling. After exploration, they completed a structured questionnaire delivered via Google Forms. The questionnaire consisted of three sections. The first applied the System Usability Scale (SUS), a widely used 10-item instrument that generates a single usability score ranging from 0 to 100 (Almasi et

al., 2023; Bangor et al., 2009). The second section included six additional user experience (UX) items rated on a 5-point Likert scale, covering layout clarity, chart readability, colour and visual appropriateness, usefulness of insights, navigation intuitiveness, and overall satisfaction. Finally, an open-ended feedback section allowed participants to comment on dashboard strengths, identify difficulties, and suggest improvements.

7.6.2 Results

A total of 14 participants completed the usability testing. The sample included both male and female respondents, with ages ranging from 24 to 42 years (mean = 28.6 years). In terms of prior experience, most participants reported at least some familiarity with dashboards and data visualization tools, ranging from slightly familiar to very familiar. This background profile indicates that the participants were representative of typical end users, such as business managers or data practitioners, who would be expected to engage with dashboards in professional settings. The System Usability Scale (SUS) produced an overall mean score of 70.8 (Standard Deviation = 16.2), with individual scores ranging from 40.0 to 97.5. According to benchmark interpretations (Bangor et al., 2009), a SUS score above 68 indicates good usability. Thus, the dashboard achieved a usability rating that falls within the "*good*" range, suggesting it is effective and usable for its intended purpose. While most participants rated the dashboard positively, the variation in scores reflects that a small subset found certain aspects more challenging. Nevertheless, the average score demonstrates that the dashboard provides a generally user-friendly interface.

Beyond the SUS, participants rated six specific user experience aspects on a five-point Likert scale. Results are summarized in Table 7.1. The results highlight navigation and layout clarity as the strongest dimensions of the dashboard, both rated above 4.0. Conversely, chart readability and colour/visual design scored slightly lower at 3.71, indicating areas that could benefit from refinement. Overall satisfaction scored 4.10, reinforcing that the majority of participants were satisfied with their interaction with the system.

Table 7.1
Summary Of User Experience Testing Result

UX Dimension	Mean	Std. Dev.	Interpretation
Layout clarity	4.14	0.65	Perceived as clear and well-structured.
Chart readability	3.71	0.85	Slightly weaker, moderate ease of interpretation.
Colours and visuals	3.71	0.96	Moderate, with room for improvement.
Usefulness for decision-making	4.05	0.80	Considered useful for practical insights.
Navigation intuitiveness	4.33	0.80	Strongest feature, highly rated.
Overall satisfaction	4.10	0.70	High level of satisfaction.

Qualitative feedback provided further insights into user perceptions. Positive comments emphasized the dashboard's clarity, ease of use, and dynamic visual presentation. Several participants appreciated that the data outputs were easy to interpret and visually engaging. On the other hand, some noted challenges with design consistency and the alignment between data and visual elements. Suggestions for improvement included adopting brighter colour schemes, improving design alignment, and in some cases, embedding AI-assisted features to further enhance interactivity. A few participants reported no difficulties and explicitly expressed satisfaction with the dashboard in its current form. Overall, the usability testing confirmed that the dashboard delivers a positive user experience. The SUS score of 70.8 places it within the good usability range, while additional UX ratings highlighted strengths in navigation and layout clarity, alongside opportunities to improve chart readability and visual design. Qualitative feedback reinforced these results, with users praising the clarity and accessibility of insights while suggesting improvements in colour schemes and design consistency. These findings align with established dashboard design principles in HCI, which emphasize balancing clarity, cognitive load management, and visual consistency (Bach et al., 2022). The strong navigation and structured layout reflect effective modular design, while weaker scores for visual aspects suggest a trade-off between minimalism and expressiveness. By addressing these areas in future iterations, the dashboard could progress from good usability toward excellence, further supporting its role as a decision-support tool for social media analytics.

7.7 Conclusion

This chapter presented the design, implementation, and evaluation of the TikTok Analysis Dashboard as the final stage of the research pipeline. Developed in Streamlit, the dashboard integrates outputs from GPT-based sentiment analysis, BERTopic-driven topic modeling, and influencer engagement analytics into a unified and interactive interface. Its modular architecture and alignment with IT design principles such as scalability, portability, and user-centred design demonstrates how complex social media data can be transformed into accessible business intelligence. The dashboard's visual modules addressed core analytical needs, including influencer revenue profiling, engagement metrics, sentiment distributions, and topic cluster exploration. Usability testing further confirmed its effectiveness, with a mean SUS score of 70.8 and positive feedback on navigation and layout clarity. While areas such as chart readability and colour design require refinement, the overall results indicate that the dashboard provides a practical, interpretable, and user-friendly platform for decision support. In summary, the dashboard functions as a methodological proof of concept for delivering AI-powered social media analytics to end users. It not only validates the research methodology but also contributes a scalable model for real-time, data-driven strategies in businesses engaging with TikTok user-generated content. The interactive dashboard is publicly accessible at: https://dashboardpy-5nqekbappvvugyvc8eaappn4.streamlit.app, enabling users to directly explore sentiment trends, topic insights, and influencer engagement metrics through the Streamlit interface.

CHAPTER 8
ACTIONABLE INSIGHTS AND CONCLUSION

8.1 Summary of Objectives and Outcomes

Table 8.1 below outlines the key research objectives and how they were achieved, along with the associated findings. Each objective was addressed through specific chapters of the research, contributing to the overall research objectives.

Table 8.1
Summary Of Research Objective's Completion

Objectives	Status	Achievement
RO1: To establish a structured and adaptable methodology for processing TikTok's unstructured data for reliable insight extraction.	Completed	Chapter 2 (Literature Review), where the research framework and methodology are defined, and Chapter 3 (Research Methodology), where the phases of the research methodology are discussed. This chapter establishes a structured and adaptable methodology for handling TikTok's unstructured data, integrating preprocessing techniques and NLP models into a modular pipeline that can be reused or extended for future research.
RO2: To analyze TikTok descriptive metadata to identify patterns, trends, and engagement factors that influence audience behavior.	Completed	Chapter 4 shows that engagement varies widely among influencers, requiring tailored strategies. Revenue depends more on engagement and smart tactics than follower count, with diverse income

		streams boosting earnings. Fewer, relevant hashtags, 35–60 second videos, and culturally or seasonally relevant content improve interaction, while emotional or valuable posts drive higher engagement. These findings highlight clear patterns, trends, and engagement drivers, directly addressing the research objective.
RO3: To apply advanced modeling techniques, including BERTopic for topic modeling and GPT-based sentiment analysis, to generate meaningful insights from TikTok comments.	Completed	Chapters 5 and 6 apply GPT-4o for sentiment analysis and BERTopic for topic modeling, then evaluate them. Manual examinations confirm the sentiment classifier at ~92% accuracy, while topic coherence scores (c_v = 0.620, c_uci = 0.918) validate the two discovered themes. The results show a mainly positive audience, but neutral, question-laden comments signal the strongest conversion potential, and negative remarks demand swift action. Topic modeling clearly separates “product-usage talk” from “sensory-feature talk”, giving brands direct guidance on messaging, packaging, and live-selling tactics. Together, these results provide actionable insights for content strategy and customer engagement.

RO4: To extract actionable business insights from TikTok data and visualise through a user-friendly dashboard for data-driven decision-making.	Completed	Chapter 7 converts the analytic outputs into a lightweight Streamlit dashboard that surfaces actionable business insights at a glance. It ranks influencers by revenue and highlights engagement factors such as video length, concise hashtag use, and posting rhythm. The dashboard lets users track sentiment shifts from positive buzz to neutral purchase-ready queries through interactive charts and visualises the two dominant content themes (product usage versus sensory features) with word clouds and sample comments. Fast to load, device-agnostic, and filterable in a few clicks, the dashboard enables data-driven decision-making by placing actionable insights directly in marketers' hands.

As summarized in Table 8.1, a systematic progression from methodological foundation to practical application. The completion of RO1 established the essential technical infrastructure, creating a robust and adaptable pipeline capable of transforming noisy, unstructured TikTok data into a clean, machine-readable format. Building upon this foundation, RO2 utilized descriptive analytics to identify key engagement factors, such as the 'Revenue Paradox', where strategic content choices outweighed traditional metrics like follower counts. The study then advanced to RO3, where sophisticated modeling through GPT-4o and BERTopic enabled the extraction of deep sentiment and thematic insights, effectively bridge the gap between raw data and consumer knowledge. Finally, RO4 culminated in the operationalization of these findings through a Streamlit dashboard, providing a user-friendly interface for data-

driven decision-making. Collectively, these achievements demonstrate that each objective served as a critical building block in converting TikTok User-Generated Content (UGC) into actionable business intelligence.

8.2 Research Contributions

This research delivers significant contributions to the domains of Business Intelligence (BI), Information Technology, and Social Media Analytics. By successfully navigating the transition from raw unstructured data to actionable wisdom, the study offers distinct theoretical, methodological, and practical advancements. The primary theoretical contribution lies in the adaptation of the DIKW (Data, Information, Knowledge, Wisdom) hierarchy and the Data Science Trajectories (DST) model specifically for the context of short-form video platforms. While previous studies have applied these frameworks to static text (like Twitter or blogs), this research demonstrates their applicability to the dynamic, unstructured, and informal nature of TikTok User-Generated Content (UGC). By mapping the DST model's iterative phases to the DIKW hierarchy, this study establishes a validated theoretical blueprint for extracting "Wisdom" (strategic insights) from "Data" (raw comments) in high-velocity digital environments. Furthermore, it enriches the academic literature on social commerce by identifying the "Revenue Paradox", theoretically challenging the traditional assumption that high positive sentiment and follower counts are linear predictors of financial success.

Methodologically, this research introduces a novel and modular Social Media Business Intelligence Methodology (SMBIM). Unlike traditional linear approaches (like CRISP-DM), this pipeline integrates advanced Generative AI (GPT-4o) for pre-processing and sentiment classification with density-based clustering (BERTopic) for thematic discovery. A key innovation is the development of a hybrid cleaning protocol that utilizes Large Language Models (LLMs) to normalize code-switching (mixed Malay/English) and interpret emoji-heavy semantics before topic modeling. This approach significantly outperforms standard rule-based NLP methods in handling the linguistic noise typical of Malaysian social media data. The construction of this reusable and end-to-end pipeline, from Apify scraping to Streamlit visualization provides a technical template for future researchers analyzing similar unstructured datasets.

For practitioners, specifically brand managers and digital marketers in the Beauty and Personal Care sector, this research offers direct commercial value. The development of the interactive Streamlit Dashboard operationalizes the research findings, transforming abstract analytics into a decision-support tool. Several key practical insights include the revelation that revenue is driven by active video sales and live streams rather than passive follower accumulation allows brands to pivot resources from "vanity metrics" to conversion-focused content. Next, the identification of specific engagement drivers such as the optimal video duration (35–60 seconds) and the efficacy of "clean" captions (0–2 hashtags) which provides an immediate tactical playbook for increasing organic reach. Lastly, the insight that "Neutral" comments often represent high-intent purchase inquiries (like "Is this safe for sensitive skin?") enables customer service teams to prioritize these leads over generic positive praise, directly impacting sales conversion.

8.3 Research Limitation

While this research has achieved its stated objectives and demonstrated the feasibility of integrating advanced natural language processing techniques into social media analytics, several limitations should be acknowledged. These limitations affect both the scope of the findings and their generalizability to broader contexts. The dataset was intentionally limited to 20 top-revenue influencers in the Beauty and Personal Care category, covering the period from January to August 2024. This choice ensured a manageable yet relevant dataset but also introduced potential biases. High-revenue influencers often represent polished and brand-driven strategies that may not reflect the practices or outcomes of mid-tier and emerging creators. Similarly, the temporal restriction risks capturing seasonal effects, such as spikes around holidays or promotional campaigns, that may not generalize across time.

For each video, only up to 100 comments were collected and analyzed. While this cap facilitated computational efficiency and streamlined preprocessing, it inherently excluded the "*long tail*" of interactions. Comments beyond the cap which often from less-engaged or late-arriving users, may contain perspectives that differ significantly from early or highly visible interactions. This limitation suggests that the sentiment distribution observed in this research may be skewed toward the most active segment of the audience. GPT-4o was selected for sentiment analysis due to its capacity

to interpret informal, emoji-rich, and multilingual text. Despite strong overall performance, challenges remain. Sarcasm, code-switching between languages, and mixed emotional tones continue to pose risks of misclassification. Minimal prompt engineering and manual spot checks improved accuracy but could not fully eliminate these risks. As a result, some nuance in audience sentiment may have been lost, particularly in comments that blended humour, irony, or cultural references. Using BERTopic, this research identified two dominant and coherent themes. While this outcome is interpretable and consistent with the dataset, it may oversimplify audience discourse. Sub-themes such as pricing concerns, product authenticity, or delivery issues may have been collapsed into broader clusters due to parameter settings in clustering and filtering. This limitation highlights the trade-off between interpretability and granularity in unsupervised modeling.

Finally, the Streamlit dashboard, while effective in translating technical outputs into accessible insights, has functional constraints. It prioritized clarity and usability over advanced interactivity. As such, its current version supports exploration of static outputs but does not incorporate real-time updates, automated alerts, or full mobile optimization. These constraints limit its applicability in fast-moving commercial environments where responsiveness and portability are essential. Overall, these limitations do not undermine the core contributions of the research but rather frame its boundaries. Recognizing them provides an opportunity to refine future work, whether by expanding datasets, incorporating multimodal features, or enhancing technical tools to capture the complexity of influencer-driven social media ecosystems.

8.4 Future Works

The limitations identified in this research provide several avenues for future research and system development. Extending this work involves not only expanding the scope of the dataset but also advancing analytical methods and technical integration to create a more robust and versatile framework for social media analytics. Future research should expand beyond the focus on top-revenue influencers in the Beauty and Personal Care sector. Including mid-tier and micro-influencers would provide a more representative picture of the ecosystem, as these creators often drive niche engagement and grassroots trends. Similarly, applying the pipeline to other domains such as fashion, gaming, or food would allow for cross-industry comparisons and test the adaptability

of the methodology. Extending the temporal range beyond the January–August 2024 window would also help capture seasonal effects and campaign-driven dynamics that shape influencer performance and audience sentiment over time.

Expanding the number of comments analyzed per video would enable the inclusion of the long-tail of user interactions, offering a more balanced view of audience perspectives. In addition, future work could place greater emphasis on multilingual and cross-cultural discourse, given the global nature of platforms like TikTok. Beyond simple sentiment classes, more nuanced labeling strategies should be considered, including aspect-based sentiment such as product quality, delivery speed, price fairness, and intent recognition such as purchase intention, complaint, and inquiry. This would allow for a more granular understanding of user feedback and its implications for influencer strategy and product positioning. While BERTopic provided clear insights in this research, more sophisticated approaches could uncover deeper patterns. Hierarchical or dynamic topic modeling, for example, would enable researchers to identify sub-themes and track how topics evolve over time. Multimodal analysis that integrates textual, visual, and audio features could also capture the full richness of TikTok content, linking video attributes such as tone, aesthetics, and audio cues to audience reactions. Furthermore, the use of causal inference methods such as panel data econometrics or quasi-experimental designs could move the analysis beyond correlation and would provide stronger evidence about the drivers of influencer revenue and engagement outcomes.

Finally, on the practical side, the Streamlit dashboard can be expanded into a more advanced decision-support tool. Potential enhancements include role-based access tailored to marketers, product managers, or influencer agencies, real-time monitoring and alert systems to flag spikes in negative or neutral sentiment, and predictive analytics modules to forecast campaign outcomes. Technical improvements should also focus on mobile optimization and lightweight deployment, enabling decision-makers to interact with insights in fast-paced environments such as live campaigns or on-the-go product launches. In summary, future work should aim to transform this proof-of-concept into a more comprehensive and adaptable system. By broadening data scope, deepening analytic techniques, and strengthening dashboard functionality, the framework has the potential to evolve into a scalable tool for both academic inquiry and industry application in influencer-driven digital ecosystems.

8.5 Conclusion

The core aim of this research is to systematically analyze TikTok user-generated content (UGC) to extract meaningful insights that can support business intelligence, with a particular focus on the Beauty and Personal Care sector in Malaysia. This research addresses the challenge of working with unstructured and bilingual TikTok data, which often contains short, informal, and slang-filled expressions that are not easily processed by traditional methods. To achieve this aim, the research develops a structured methodology by adapting the Data Science Trajectories (DST) model, which provides a flexible yet rigorous framework for handling the full pipeline of social media data analysis, from acquisition to actionable outputs. By applying this methodology, the research seeks to transform raw TikTok comments into structured, decision-ready insights. Specifically, the research employs advanced natural language processing (NLP) techniques, including transformer-based embeddings and BERTopic, to perform topic modeling that identifies recurring themes in consumer discussions. In parallel, GPT-based sentiment analysis is applied to classify comments into positive, neutral, or negative categories, while also accounting for mixed expressions, emojis, and Malay slang. This dual approach allows for a deeper understanding of both what consumers are discussing and how they feel about it. The methodological pipeline is not limited to theoretical contributions, it extends to mapping the extracted insights into a Streamlit-based dashboard that visualizes sentiment trends, influencer engagement metrics, and thematic patterns. The dashboard's usability was validated through a questionnaire of System Usability Scale test (mean 70.8, rated "*good*"), with strong scores for navigation and layout, supporting its role as a decision-support tool.

The research's central goal is not only to evaluate model performance in handling informal TikTok text but also to generate insights that can guide brands and influencers in making data-driven decisions. The analysis highlights patterns such as the role of neutral sentiment in consumer inquiries, the impact of positive versus negative sentiment on engagement, and the thematic emphasis consumers place on product features and usage experiences. These findings demonstrate how structured analytics can provide businesses with actionable intelligence insights and revealing opportunities for more targeted marketing strategies and improved influencer collaborations. Ultimately, this research contributes both conceptually and practically to the field of social media business intelligence. Conceptually, it demonstrates how the

DIKW and DST framework can be adapted for fast-moving and informal UGC. Technically, it introduces a modular NLP pipeline that integrates GPT-based preprocessing, sentiment classification, BERTopic for clustering, and dashboard visualization. From an applied perspective, it shows that structured analysis of TikTok data can reveal consumer sentiment trends, engagement drivers, and thematic insights that directly support strategic decision-making. In doing so, the research bridges methodological innovation with practical application, reflecting the core aim outlined in Chapter 1 which to harness TikTok UGC as a valuable source of business intelligence for understanding consumer behavior and shaping effective marketing strategies.

REFERENCES

-, O., -, M., Anggraini, W., -, S., & Pranggono, B. (2022). Assessing Digital Readiness of Small Medium Enterprises: Intelligent Dashboard Decision Support System. *International Journal of Advanced Computer Science and Applications*, *13*(4). https://doi.org/10.14569/IJACSA.2022.0130412

Abasova, J., Tanuska, P., & Rydzi, S. (2021). Big data—knowledge discovery in production industry data storages—implementation of best practices. *Applied Sciences (Switzerland)*, *11*(16). https://doi.org/10.3390/app11167648

Abdullah, M. F., Mohamad Khan, N. R., Ibrahim, M. A., & Putit, L. (2023). Exploring the Influence of Social Media Influencers' (SMIs) Traits on Consumer Purchasing Behavior for Online Products on the TikTok Platform: The Mediating Effect of Trustworthiness. *International Journal of Academic Research in Business and Social Sciences*, *13*(11). https://doi.org/10.6007/IJARBSS/v13-i11/19605

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9.

Adanyin, A. (2024). *Ethical AI in Retail: Consumer Privacy and Fairness*.

Ahmad Asmawi, M. A. H., & Isawasan, P. (2024). *Top 20 TikTok Beauty & Personal Care Influencers*. Kaggle. https://doi.org/10.34740/KAGGLE/DS/6111672

Ali Hakami, N., & Hosni Mahmoud, H. A. (2022). The Prediction of Consumer Behavior from Social Media Activities. *Behavioral Sciences*, *12*(8), 284. https://doi.org/10.3390/bs12080284

Ali, I., Balta, M., & Papadopoulos, T. (2023). Social media platforms and social enterprise: Bibliometric analysis and systematic review. *International Journal of Information Management*, *69*, 102510. https://doi.org/10.1016/j.ijinfomgt.2022.102510

Ali, T., Omar, B., & Soulaimane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. *MethodsX*, *9*, 101894. https://doi.org/10.1016/j.mex.2022.101894

Almadani, B., Kaisar, H., Thoker, I. R., & Aliyu, F. (2025). A Systematic Survey of Distributed Decision Support Systems in Healthcare. *Systems*, *13*(3), 157. https://doi.org/10.3390/systems13030157

Almasi, S., Bahaadinbeigy, K., Ahmadi, H., Sohrabei, S., & Rabiei, R. (2023). Usability Evaluation of Dashboards: A Systematic Literature Review of Tools. *BioMed*

Research International, *2023*(1). https://doi.org/10.1155/2023/9990933

Almeida, M. D., Maia, V. M., Tommasetti, R., & Leite, R. de O. (2021). Sentiment analysis based on a social media customised dictionary. *MethodsX*, *8*. https://doi.org/10.1016/j.mex.2021.101449

Al-Okaily, A., Teoh, A. P., & Al-Okaily, M. (2023). Evaluation of data analytics-oriented business intelligence technology effectiveness: an enterprise-level analysis. *Business Process Management Journal*, *29*(3), 777–800. https://doi.org/10.1108/BPMJ-10-2022-0546

Amiri, A. M., Kushwaha, B. P., & Singh, R. (2023). Visualisation of global research trends and future research directions of digital marketing in small and medium enterprises using bibliometric analysis. *Journal of Small Business and Enterprise Development*, *30*(3), 621–641. https://doi.org/10.1108/JSBED-04-2022-0206

Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives. *Applied Sciences*, *13*(6), 3911. https://doi.org/10.3390/app13063911

Aramburu, M. J., Berlanga, R., & Lanza-Cruz, I. (2023). A Data Quality Multidimensional Model for Social Media Analysis. *Business & Information Systems Engineering*. https://doi.org/10.1007/s12599-023-00840-9

Bach, B., Freeman, E., Abdul-Rahman, A., Turkay, C., Khan, S., Fan, Y., & Chen, M. (2022). Dashboard Design Patterns. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. https://doi.org/10.1109/TVCG.2022.3209448

Bahtar, A. Z., & Muda, M. (2016). The Impact of User – Generated Content (UGC) on Product Reviews towards Online Purchasing – A Conceptual Framework. *Procedia Economics and Finance*, *37*, 337–342. https://doi.org/10.1016/S2212-5671(16)30134-4

Balakrishnan, V., Ng, K. S., & Rahim, H. A. (2021). To share or not to share – The underlying motives of sharing fake news amidst the COVID-19 pandemic in Malaysia. *Technology in Society*, *66*, 101676. https://doi.org/10.1016/j.techsoc.2021.101676

Bangor, A., Kortum, P., & Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, *4*(3), 114–123.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App).

Journal of Business and Psychology. https://doi.org/10.15139/S3/R4W7ZS

Baskarada, S., & Koronios, A. (2013). Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension. *Australasian Journal of Information Systems*, *18*(1). https://doi.org/10.3127/ajis.v18i1.748

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI and Society*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4

Bauer, P. C., & Clemm von Hohenberg, B. (2021). Believing and Sharing Information by Fake Sources: An Experiment. *Political Communication*, *38*(6), 647–671. https://doi.org/10.1080/10584609.2020.1840462

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. http://arxiv.org/abs/1903.10676

Bhandari, A., & Bimo, S. (2020). TIKTOK AND THE "ALGORITHMIZED SELF": A NEW MODEL OF ONLINE INTERACTION. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11172

Bharadiya, J. P. (2023). The role of machine learning in transforming business intelligence. *International Journal of Computing and Artificial Intelligence*, *4*(1), 16–24. https://doi.org/10.33545/27076571.2023.v4.i1a.60

Bianchi, F., Terragni, S., & Hovy, D. (2020). *Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence*. http://arxiv.org/abs/2004.03974

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).

Bossetta, M. (2018). The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism and Mass Communication Quarterly*, *95*(2), 471–496. https://doi.org/10.1177/1077699018763307

Brooke, J. (2013). SUS: A Retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Camilleri, M. A. (2024). Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Systems*, *41*(7). https://doi.org/10.1111/exsy.13406

Campagnolo, J. M., Duarte, D., & Dal Bianco, G. (2022). Topic Coherence Metrics: How Sensitive Are They? *Journal of Information and Data Management*.

Chan, M. (2024). Verification Behaviors and Countermeasures in the Age of Misinformation. *Journalism and Mass Communication Quarterly*, *101*(1), 13–19. https://doi.org/10.1177/10776990231223998

Cheng, Z., & Li, Y. (2024). Like, Comment, and Share on TikTok: Exploring the Effect of Sentiment and Second-Person View on the User Engagement with TikTok News Videos. *Social Science Computer Review*, *42*(1), 201–223. https://doi.org/10.1177/08944393231178603

Cho, B. (2024). A Sentiment Analysis to investigate the sentiment of users on fad diet content on Tik Tok. In *Journal of High School Science* (Vol. 8, Issue 4).

Cho, H., Cannon, J., Lopez, R., & Li, W. (2024). Social media literacy: A conceptual framework. *New Media & Society*, *26*(2), 941–960. https://doi.org/10.1177/14614448211068530

Chowdhury, I., Moeid, A., Hoque, E., Kabir, M. A., Hossain, Md. S., & Islam, M. M. (2021). Designing and Evaluating Multimodal Interactions for Facilitating Visual Analysis With Dashboards. *IEEE Access*, *9*, 60–71. https://doi.org/10.1109/ACCESS.2020.3046623

Corallo, A., Errico, F., Fortunato, L., Spennato, A., & De Blasi, C. (2024). Effects Influence of Social Media Constructs on Shopping: An Empirical Study on the Prediction of Retail Clothing Sales. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-024-01827-x

Cuomo, M. T., Tortora, D., Giordano, A., Festa, G., Metallo, G., & Martinelli, E. (2020). User-generated content in the era of digital well-being: A netnographic analysis in a healthcare marketing context. *Psychology & Marketing*, *37*(4), 578–587. https://doi.org/10.1002/mar.21327

Darmatama, M., & Erdiansyah, R. (2021). *The Influence of Advertising in Tiktok Social Media and Beauty Product Image on Consumer Purchase Decisions*.

de Mast, J., & Lokkerbol, J. (2024). DAPS diagrams for defining Data Science projects. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00916-7

Delmonaco, D., Mayworm, S., Thach, H., Guberman, J., Augusta, A., & Haimson, O. L. (2024). "What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1). https://doi.org/10.1145/3637431

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Di Minin, E., Fink, C., Hausmann, A., Kremer, J., & Kulkarni, R. (2021). How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*, *35*(2), 437–446. https://doi.org/10.1111/cobi.13708

Djafarova, E., & Rushworth, C. (2017). Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users. *Computers in Human Behavior*, *68*, 1–7. https://doi.org/10.1016/j.chb.2016.11.009

Dolatabadi, S. H., Gatial, E., Budinská, I., & Balogh, Z. (2024). Integrating Human-Computer Interaction Principles in User-Centered Dashboard Design: Insights from Maintenance Management. *2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES)*, 000219–000224. https://doi.org/10.1109/INES63318.2024.10629098

Donoho, D. (2017). 50 Years of Data Science. In *Journal of Computational and Graphical Statistics* (Vol. 26, Issue 4, pp. 745–766). American Statistical Association. https://doi.org/10.1080/10618600.2017.1384734

Donohoe, D., & Costello, E. (2020). Data Visualisation Literacy in Higher Education: An Exploratory Study of Understanding of a Learning Dashboard Tool. *International Journal of Emerging Technologies in Learning (IJET)*, *15*(17), 115. https://doi.org/10.3991/ijet.v15i17.15041

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Elmitwalli, S., & Mehegan, J. (2024). Sentiment analysis of COP9-related tweets: a comparative study of pre-trained models and traditional techniques. *Frontiers in Big Data*, *7*. https://doi.org/10.3389/fdata.2024.1357926

Emmanuel Osamuyimen Eboigbe, Oluwatoyin Ajoke Farayola, Funmilola Olatundun Olatoye, Obiageli Chinwe Nnabugwu, & Chibuike Daraojimba. (2023). BUSINESS INTELLIGENCE TRANSFORMATION THROUGH AI AND DATA ANALYTICS. *Engineering Science & Technology Journal*, *4*(5), 285–307. https://doi.org/10.51594/estj.v4i5.616

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74–81. https://doi.org/10.1145/2602574

Feroz Khan, G., & Vong, S. (2014). Virality over YouTube: an empirical analysis.

Internet Research, *24*(5), 629–647. https://doi.org/10.1108/IntR-05-2013-0085

Frazao, D. A. G., Costa, T. S. A. da, Araujo, T. D. O. de, Meiguins, B. S., & Santos, C. G. R. dos. (2021). A brief review of dashboard visualizations employed to support management or business decisions. *2021 25th International Conference Information Visualisation (IV)*, 100–107. https://doi.org/10.1109/IV53921.2021.00025

Gandhi, U. D., Malarvizhi Kumar, P., Chandra Babu, G., & Karthick, G. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). *Wireless Personal Communications*. https://doi.org/10.1007/s11277-021-08580-3

Gesmundo, M. A. G., Jordan, M. D. S., Meridor, W. H. D., Muyot, D. V., Castano, M. C. N., & Bandojo, A. J. P. (2022). TikTok as a Platform for Marketing Campaigns: The effect of Brand Awareness and Brand Recall on the Purchase Intentions of Millennials. *Journal of Business and Management Studies*, *4*(2), 343–361. https://doi.org/10.32996/jbms.2022.4.2.27

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). *Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey*. http://arxiv.org/abs/2109.02508

González-Bailón, S., & Lelkes, Y. (2023). Do social media undermine social cohesion? A critical review. *Social Issues and Policy Review*, *17*(1), 155–180. https://doi.org/10.1111/sipr.12091

Graser, A., Jalali, A., Lampert, J., Weißenfeld, A., & Janowicz, K. (2024). MobilityDL: a review of deep learning from trajectory data. *GeoInformatica*. https://doi.org/10.1007/s10707-024-00518-8

Grieves, M. (2024). DIKW as a General and Digital Twin Action Framework: Data, Information, Knowledge, and Wisdom. *Knowledge*, *4*(2), 120–140. https://doi.org/10.3390/knowledge4020007

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. http://arxiv.org/abs/2203.05794

Grossmann, W., & Rinderle-Ma, S. (2015). *Fundamentals of Business Intelligence*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-46531-8

Gupta, S., Singh, A., & Kumar, V. (2023). Emoji, Text, and Sentiment Polarity Detection Using Natural Language Processing. *Information*, *14*(4), 222. https://doi.org/10.3390/info14040222

Gutierrez, A., Punjaisri, K., Desai, B., Syed Alwi, S. F., O'Leary, S., Chaiyasoonthorn, W., & Chaveesuk, S. (2023). Retailers, don't ignore me on social media! The importance of consumer-brand interactions in raising purchase intention - Privacy the Achilles heel. *Journal of Retailing and Consumer Services*, *72*, 103272. https://doi.org/10.1016/j.jretconser.2023.103272

Hadjimichael, A., Schlumberger, J., & Haasnoot, M. (2024). Data visualisation for decision making under deep uncertainty: current challenges and opportunities. *Environmental Research Letters*, *19*(11), 111011. https://doi.org/10.1088/1748-9326/ad858b

Hamzehi, M., & Hosseini, S. (2022). Business intelligence using machine learning algorithms. *Multimedia Tools and Applications*, *81*(23), 33233–33251. https://doi.org/10.1007/s11042-022-13132-3

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Harriger, J. A., Thompson, J. K., & Tiggemann, M. (2023). TikTok, TikTok, the time is now: Future directions in social media and body image. *Body Image*, *44*, 222–226. https://doi.org/10.1016/j.bodyim.2023.01.005

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Hmoud, H., Al-Adwan, A. S., Horani, O., Yaseen, H., & Zoubi, J. Z. Al. (2023). Factors influencing business intelligence adoption by higher education institutions. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*(3), 100111. https://doi.org/10.1016/j.joitmc.2023.100111

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for*

Sentiment Analysis of Social Media Text. http://sentic.net/

Islam, M. R., Akter, S., Islam, L., Razzak, I., Wang, X., & Xu, G. (2024). Strategies for evaluating visual analytics systems: A systematic review and new perspectives. *Information Visualization*, *23*(1), 84–101. https://doi.org/10.1177/14738716231212568

Issa, T., & Isaias, P. (2022). Usability and Human–Computer Interaction (HCI). In *Sustainable Design* (pp. 23–40). Springer London. https://doi.org/10.1007/978-1-4471-7513-1_2

Jahani, H., Jain, R., & Ivanov, D. (2023). Data science and big data analytics: a systematic review of methodologies used in the supply chain and logistics research. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05390-7

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and Challenges of Big Data Research. *Big Data Research*, *2*(2), 59–64. https://doi.org/10.1016/j.bdr.2015.01.006

Johnson, N. E., Short, J. C., Chandler, J. A., & Jordan, S. L. (2022). Introducing the contentpreneur: Making the case for research on content creation-based online platforms. *Journal of Business Venturing Insights*, *18*. https://doi.org/10.1016/j.jbvi.2022.e00328

Joseph, T. T., Wax, D. B., Goldstein, R., Huang, J., McCormick, P. J., & Levin, M. A. (2020). A Web-Based Perioperative Dashboard as a Platform for Anesthesia Informatics Innovation. *Anesthesia & Analgesia*, *131*(5), 1640–1645. https://doi.org/10.1213/ANE.0000000000005193

Kanan, T., Mughaid, A., Al-Shalabi, R., Al-Ayyoub, M., Elbes, M., & Sadaqa, O. (2023). Business intelligence using deep learning techniques for social media contents. *Cluster Computing*, *26*(2), 1285–1296. https://doi.org/10.1007/s10586-022-03626-y

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Ke, J., Liao, P., Li, J., & Luo, X. (2023). Effect of information load and cognitive style on cognitive load of visualized dashboards for construction-related activities. *Automation in Construction*, *154*, 105029. https://doi.org/10.1016/j.autcon.2023.105029

Kemp, S. (2024). *Digital 2024: Global Overview Report*.

Kewsuwun, N., & Kajornkasirat, S. (2022). A sentiment analysis model of Agritech startup on Facebook comments using naive Bayes classifier. *International Journal of Electrical and Computer Engineering (IJECE)*, *12*(3), 2829. https://doi.org/10.11591/ijece.v12i3.pp2829-2838

Kheiri, K., & Karimi, H. (2023). *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning*.

Kim, P. W. (2024). A Framework to Overcome the Dark Side of Generative Artificial Intelligence (GAI) Like ChatGPT in Social Media and Education. *IEEE Transactions on Computational Social Systems*, *11*(4), 5266–5274. https://doi.org/10.1109/TCSS.2023.3315237

Kong, Q., Mao, W., Chen, G., & Zeng, D. (2020). Exploring Trends and Patterns of Popularity Stage Evolution in Social Media. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(10), 3817–3827. https://doi.org/10.1109/TSMC.2018.2855806

Lanza-Cruz, I., Berlanga, R., & Aramburu, M. J. (2024). Multidimensional Author Profiling for Social Business Intelligence. *Information Systems Frontiers*, *26*(1), 195–215. https://doi.org/10.1007/s10796-023-10370-0

Lee, J., Cameron, I., & Hassall, M. (2022). Information needs and challenges in future process safety. *Digital Chemical Engineering*, *3*, 100017. https://doi.org/10.1016/j.dche.2022.100017

Li, Z. (2022). Strategies Behind Tik Tok's Global Rise. *Proceedings of the 2022 International Conference on Social Sciences and Humanities and Arts*.

Lim, W. M., & Kumar, S. (2024). Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach. *Global Business and Organizational Excellence*, *43*(2), 17–26. https://doi.org/10.1002/joe.22229

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. http://arxiv.org/abs/1907.11692

Lo Duca, A., & McDowell, K. (2024). Using the S-DIKW framework to transform data visualization into data storytelling. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24973

Macías, J. A., & Borges, C. R. (2024). Monitoring and forecasting usability indicators: A business intelligence approach for leveraging user-centered evaluation data.

Science of Computer Programming, *234*, 103077. https://doi.org/10.1016/j.scico.2023.103077

Maia, S., Teixeira Domingues, J. P., Rocha Varela, M. L. R., & Fonseca, L. M. (2024). Exploring the user-generated content data to improve quality management. *The TQM Journal*. https://doi.org/10.1108/TQM-09-2023-0278

Malzer, C., & Baum, M. (2019). *A Hybrid Approach To Hierarchical Density-based Cluster Selection*. https://doi.org/10.1109/MFI49285.2020.9235263

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Marti-Ochoa, J., Martin-Fuentes, E., & Ferrer-Rosell, B. (2024). Airbnb on TikTok: Brand Perception Through User Engagement and Sentiment Trends. *Social Science Computer Review*. https://doi.org/10.1177/08944393241260242

Mekala, S., Padmaja Rani, B., & of CSE, P. (2019). *Kernel PCA Based Dimensionality Reduction Techniques for preprocessing of Telugu text documents for Cluster Analysis*. www.ijert.org

Monacho, B. C., & Slamet, Y. (2023). The Effect of Influencer Engagement Rate in Increasing Followers of Instagram Official Account. *Jurnal Komunikasi: Malaysian Journal of Communication*, *39*(2), 373–388. https://doi.org/10.17576/JKMJC-2023-3902-21

Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., & Daudpota, S. M. (2024). Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis. *IEEE Access*, *12*, 60943–60959. https://doi.org/10.1109/ACCESS.2024.3386969

Naab, T. K., & Sehl, A. (2017). Studies of user-generated content: A systematic review. *Journalism*, *18*(10), 1256–1273. https://doi.org/10.1177/1464884916673557

Nanda, P., & Kumar, V. (2021). Social Media Analytics: Tools, Techniques and Present Day Practices. *International Journal of Services Operations and Informatics*, *11*(4), 1. https://doi.org/10.1504/IJSOI.2021.10039351

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). *BERTweet: A pre-trained language model for English Tweets*. http://arxiv.org/abs/2005.10200

Nielsen, J. (1994, April 24). *Usability Inspection Methods*.

Öztürk, O., Kocaman, R., & Kanbach, D. K. (2024). How to design bibliometric research: an overview and a framework proposal. *Review of Managerial Science*, *18*(11), 3333–3361. https://doi.org/10.1007/s11846-024-00738-0

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques*. EMNLP. http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Peng, H., Mengni, Z., Yang, L., Juanatas, R., Niguidula, J., & Huiliang, H. (2023). Research on Brand Marketing Strategy on Tik Tok Short Video Platform. *SHS Web of Conferences*, *159*, 02024. https://doi.org/10.1051/shsconf/202315902024

Pereira, B. B., & Ha, S. (2024). ENVIRONMENTAL ISSUES ON TIKTOK: TOPICS AND CLAIMS OF MISLEADING INFORMATION. *Journal of Baltic Science Education*, *23*(1), 131–150. https://doi.org/10.33225/jbse/24.23.131

Peters, M. A., Jandrić, P., & Green, B. J. (2024). The DIKW Model in the Age of Artificial Intelligence. *Postdigital Science and Education*. https://doi.org/10.1007/s42438-024-00462-8

Praful Bharadiya, J. (2023). A Comparative Study of Business Intelligence and Artificial Intelligence with Big Data Analytics. *American Journal of Artificial Intelligence*. https://doi.org/10.11648/j.ajai.20230701.14

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications*, *9*(1), 12. https://doi.org/10.3390/publications9010012

Puchakayala, S. J., T., A. B., Ravikumar, A., & Sriraman, H. (2025). Nomad Analytix: Text-rich visual reasoning using vision models for insights and recommendations. *Software Impacts*, *24*, 100765. https://doi.org/10.1016/j.simpa.2025.100765

Rahm, E., & Hong, H. Do. (2000). Data Cleaning: Problem and Current Approaches. *IEEE Computer Society*, *23*. http://list.research.microsoft.com/scripts/lyris.pl?enter=debull.

Ravichandran, P., Machireddy, J. R., & Rachakatla, S. K. (2022). AI-Enhanced Data Analytics for Real Time Business Intelligence: Applications and Challenges. *Journal of AI in Healthcare and Medicine By Health Science Publishers International*, *2*(2).

Ray, P., & Chakrabarti, A. (2022). A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis. *Applied Computing and Informatics*, *18*(1/2), 163–178.

https://doi.org/10.1016/j.aci.2019.02.002

Rehman, A., Naz, S., & Razzak, I. (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, *28*(4), 1339–1371. https://doi.org/10.1007/s00530-020-00736-8

Richter Lagha, R., Burningham, Z., Sauer, B. C., Leng, J., Peters, C., Huynh, T., Patel, S., Halwani, A. S., & Kramer, B. J. (2020). Usability Testing a Potentially Inappropriate Medication Dashboard: A Core Component of the Dashboard Development Process. *Applied Clinical Informatics*, *11*(04), 528–534. https://doi.org/10.1055/s-0040-1714693

Rippa, P., & Secundo, G. (2019). Digital academic entrepreneurship: The potential of digital technologies on academic entrepreneurship. *Technological Forecasting and Social Change*, *146*, 900–911. https://doi.org/10.1016/j.techfore.2018.07.013

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. In *Expert Systems with Applications* (Vol. 223). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2023.119862

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180. https://doi.org/10.1177/0165551506070706

Ruoff, M., Gnewuch, U., Maedche, A., & Scheibehenne, B. (2023). Designing Conversational Dashboards for Effective Use in Crisis Response. *Journal of the Association for Information Systems*, *24*(6), 1500–1526. https://doi.org/10.17705/1jais.00801

Ryu, Y. S., Yost, B., Convertino, G., Chen, J., & North, C. (2003). Exploring Cognitive Strategies for Integrating Multiple-View Visualizations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(3), 591–595. https://doi.org/10.1177/154193120304700371

Sanchez-Nunez, P., Cobo, M. J., Heras-Pedrosa, C. D. Las, Pelaez, J. I., & Herrera-Viedma, E. (2020). Opinion Mining, Sentiment Analysis and Emotion Understanding in Advertising: A Bibliometric Analysis. *IEEE Access*, *8*, 134563–

134576. https://doi.org/10.1109/ACCESS.2020.3009482

Santos, M. L. B. dos. (2022). The “so-called” UGC: an updated definition of user-generated content in the age of social media. In *Online Information Review* (Vol. 46, Issue 1, pp. 95–113). Emerald Group Holdings Ltd. https://doi.org/10.1108/OIR-06-2020-0258

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. In *SN Computer Science* (Vol. 2, Issue 5). Springer. https://doi.org/10.1007/s42979-021-00765-8

Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2023). Privacy concerns in social media UGC communities: Understanding user behavior sentiments in complex networks. *Information Systems and E-Business Management*. https://doi.org/10.1007/s10257-023-00631-5

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Scoccola, L., & Rolle, A. (2023a). Persistable: persistent and stable clustering. *Journal of Open Source Software*, *8*(83), 5022. https://doi.org/10.21105/joss.05022

Scoccola, L., & Rolle, A. (2023b). Persistable: persistent and stable clustering. *Journal of Open Source Software*, *8*(83), 5022. https://doi.org/10.21105/joss.05022

Shabrina, A. N., Sugiana, D., Dewi, Y., Sunarya, R., Studi, P., & Komunikasi, M. (2024). Pengaruh Kredibilitas Host Live Streaming Tiktok terhadap Proses Keputusan Pembelian pada Penonton Live Streaming @Skintific_Id. *Jurnal Pendidikan Tambusai*, *8*, 15383–15395.

Shahadat Hosen, M., Islam, R., Naeem, Z., Folorunso, E. O., Chu, T. S., Al Mamun, A., & Orunbon, N. O. (2024). Data-Driven Decision Making: Advanced Database Systems for Business Intelligence. *Nanotechnology Perceptions*, *20*(S3), 687–704. https://doi.org/10.62441/nano-ntp.v20iS3.51

Shaikh, A. R., Koop, D., Alhoori, H., & Sun, M. (2022). Toward Systematic Design Considerations of Organizing Multiple Views. *2022 IEEE Visualization and Visual Analytics (VIS)*, 105–109. https://doi.org/10.1109/VIS54862.2022.00030

Shanbhag, A., Jadhav, S., Thakurdesai, A., Sinare, R., & Joshi, R. (2025). *Non-*

Contextual BERT or FastText? A Comparative Analysis. http://arxiv.org/abs/2411.17661

Shien, O. Y., Huei, N. S., & Yan, N. L. (2023). The Impact of Social Media Marketing on Young Consumers' Purchase Intention in Malaysia: The Mediating Role of Consumer Engagement. *International Journal of Academic Research in Business and Social Sciences*, *13*(1). https://doi.org/10.6007/IJARBSS/v13-i1/15806

Shutsko, A. (2020). User-generated short video content in social media. a case study of tiktok. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12195 LNCS*, 108–125. https://doi.org/10.1007/978-3-030-49576-3_8

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/W14-3110

Singh, S., & Sai Vijay, T. (2024). Technology roadmapping for the e-commerce sector: A text-mining approach. *Journal of Retailing and Consumer Services*, *81*, 103977. https://doi.org/10.1016/j.jretconser.2024.103977

Singh, S., Singh, S., Koohang, A., Sharma, A., & Dhir, S. (2023). Soft computing in business: exploring current research and outlining future research directions. *Industrial Management & Data Systems*, *123*(8), 2079–2127. https://doi.org/10.1108/IMDS-02-2023-0126

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

Skarpathiotaki, C. G., & Psannis, K. E. (2022). Cross-Industry Process Standardization for Text Analytics. *Big Data Research*, *27*, 100274. https://doi.org/10.1016/j.bdr.2021.100274

Steinert, S., Marin, L., & Roeser, S. (2025). Feeling and thinking on social media: emotions, affective scaffolding, and critical thinking. *Inquiry*, *68*(1), 114–141. https://doi.org/10.1080/0020174X.2022.2126148

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Syed, S., & Spruit, M. (2017). Full-Text or abstract? Examining topic coherence scores

using latent dirichlet allocation. *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, *2018-January*, 165–174. https://doi.org/10.1109/DSAA.2017.61

Talafidaryani, M., Jalali, S. M. J., & Moro, S. (2023). Tracing the evolution of digitalisation research in business and management fields: Bibliometric analysis, topic modelling and deep learning trend forecasting. *Journal of Information Science*. https://doi.org/10.1177/01655515221148365

Tariq, R., Isawasan, P., Shamugam, L., Ahmad Asmawi, M. A. H., Nor Azman, N. A., & Zolkepli, I. A. (2025). Exploring Social Media Research Trends in Malaysia using Bibliometric Analysis and Topic Modelling. *ICST Transactions on Scalable Information Systems*, *12*. https://doi.org/10.4108/eetsis.7003

Tavoschi, L., Quattrone, F., D'Andrea, E., Ducange, P., Vabanesi, M., Marcelloni, F., & Lopalco, P. L. (2020). Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines & Immunotherapeutics*, *16*(5), 1062–1069. https://doi.org/10.1080/21645515.2020.1714311

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39. https://doi.org/10.1177/875647939000600106

Thorgren, E., Mohammadinodooshan, A., & Carlsson, N. (2024). Temporal Dynamics of User Engagement on Instagram: A Comparative Analysis of Album, Photo, and Video Interactions. *ACM Web Science Conference*, 224–234. https://doi.org/10.1145/3614419.3644029

Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing. In *Frontiers in Artificial Intelligence* (Vol. 4). Frontiers Media S.A. https://doi.org/10.3389/frai.2021.576892

Tsiu, S., Ngobeni, M., Mathabela, L., & Thango, B. (2024). *Applications and Competitive Advantages of Data Mining and Business Intelligence in SMEs Performance: A Systematic Review*. https://doi.org/10.20944/preprints202409.0940.v1

Tuğral, A., Eliiyi, U., Özdemir, K., Ergin, G., & Bakar, Y. (2021). A NEW ERA OF SEEKING KNOWLEDGE FOR #LYMPHEDEMA ON SOCIAL MEDIA: A

DETAILED INSTAGRAM HASHTAG ANALYSIS. *Lymphology*, *54*(2), 68–77. https://doi.org/10.2458/LYMPH.4728

Uma Maheswari, S., & Dhenakaran, S. S. (2023). Opinion Mining on Integrated Social Networks and E-Commerce Blog. *IETE Journal of Research*, *69*(4), 2080–2088. https://doi.org/10.1080/03772063.2021.1886603

Vijayarani, S., & Research Scholar, M. P. (2015). *Preprocessing Techniques for Text Mining-An Overview*.

Vlase, I., & Lähdesmäki, T. (2023). A bibliometric analysis of cultural heritage research in the humanities: The Web of Science as a tool of knowledge management. *Humanities and Social Sciences Communications*, *10*(1). https://doi.org/10.1057/s41599-023-01582-5

Wang, Z., Chen, J., Chen, J., & Chen, H. (2024). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, *129*(11), 7359–7384. https://doi.org/10.1007/s11192-023-04776-5

Wardle, C., & Williams, A. (2010). Beyond user-generated content: a production study examining the ways in which UGC is used at the BBC. *Media, Culture & Society*, *32*(5), 781–799. https://doi.org/10.1177/0163443710373953

Werner, J., Beisswanger, P., Schürger, C., Klaiber, M., & Theissler, A. (2023). From Data to Wisdom: A Review of Applications and Data Value in the context of Small Data. *Procedia Computer Science*, *225*, 1251–1260. https://doi.org/10.1016/j.procs.2023.10.113

Weyant, E. (2022). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th Edition. *Journal of Electronic Resources in Medical Libraries*, *19*(1–2), 54–55. https://doi.org/10.1080/15424065.2022.2046231

Xu, M., Ng, W. C., Lim, W. Y. B., Kang, J., Xiong, Z., Niyato, D., Yang, Q., Shen, X. S., & Miao, C. (2022). *A Full Dive into Realizing the Edge-enabled Metaverse: Visions, Enabling Technologies,and Challenges*. http://arxiv.org/abs/2203.05471

Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, *3*, 100073. https://doi.org/10.1016/j.dajour.2022.100073

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management*, *37*(6), 673–683. https://doi.org/10.1016/j.ijinfomgt.2017.06.004

Xu, Z., Vail, C., Kohli, A. S., & Tajdini, S. (2021). Understanding changes in a brand's core positioning and customer engagement: a sentiment analysis of a brand-owned Facebook site. *Journal of Marketing Analytics*, *9*(1), 3–16. https://doi.org/10.1057/s41270-020-00099-z

Yamagishi, K., Canayong, D., Domingo, M., Maneja, K. N., Montolo, A., & Siton, A. (2024). User-generated content on Gen Z tourist visit intention: a stimulus-organism-response approach. *Journal of Hospitality and Tourism Insights*, *7*(4), 1949–1973. https://doi.org/10.1108/JHTI-02-2023-0091

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, *27*(2), 634–639. https://doi.org/10.1016/j.chb.2010.04.014

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer) - A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497

Yew, R. L. H., Suhaidi, S. B., Seewoochurn, P., & Sevamalai, V. K. (2018). Social Network Influencers' Engagement Rate Algorithm Using Instagram Data. *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, 1–8. https://doi.org/10.1109/ICACCAF.2018.8776755

Yuli Wijayanti, Ahmad Tri Hidayat, & Intan Puspitasari. (2024). PENGARUH KREDIBILITAS SOSIAL MEDIA INFLUENCER, HUBUNGAN PARASOSIAL DAN BRAND IMAGE TERHADAP NIAT BELI KONSUMEN PADA PENGGUNA TIKTOK. *Jurnal Riset Ekonomi*, *4*.

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social Media Analytics and Intelligence. *IEEE Computer Society*, *10*, 1541–1672. http://ComputingNow.computer.org.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). *Sentiment Analysis in the Era of Large Language Models: A Reality Check*. http://arxiv.org/abs/2305.15005

Zhang, W., Yang, Y., & Liang, H. (2023). A Bibliometric Analysis of Enterprise Social Media in Digital Economy: Research Hotspots and Trends. *Sustainability*, *15*(16), 12545. https://doi.org/10.3390/su151612545

Zhang, Y., Wu, M., Tian, G. Y., Zhang, G., & Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. *Knowledge-Based Systems*, *222*, 106994. https://doi.org/10.1016/j.knosys.2021.106994

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, *29*(4), 694–700. https://doi.org/10.1016/j.ijhm.2010.02.002

Zhao, X., Wang, M., Zhao, X., Li, J., Zhou, S., Yin, D., Li, Q., Tang, J., & Guo, R. (2023). *Embedding in Recommender Systems: A Survey*. http://arxiv.org/abs/2310.18608

Zsila, Á., & Reyes, M. E. S. (2023). Pros & cons: impacts of social media on mental health. *BMC Psychology*, *11*(1), 201. https://doi.org/10.1186/s40359-023-01243-x

APPENDICES

APPENDIX 1

Testing Dashboard Usability Testing Questionnaire.

This questionnaire is designed to evaluate the usability and user experience of the developed dashboard. All responses are anonymous and will be used for academic research purposes only.

Section 1	Demographics. This section gathers basic information about the participants to contextualize the usability results.	
No.	Questions	Description
1	Age	Open-ended field
2	Gender	Male or Female
3	Familiarity with dashboards / data visualization tools	Not Familiar at all, Slightly Familiar, Moderately Familiar, Very Familiar, Expert

Section 2	System Usability Scale (SUS). Please rate the following statements on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).
No.	Questions
1	I think that I would like to use this dashboard frequently.
2	I found the dashboard unnecessarily complex.
3	I thought the dashboard was easy to use.
4	I think that I would need the support of a technical person to use this dashboard.
5	I found the various functions in this dashboard were well integrated.
6	I thought there was too much inconsistency in this dashboard.
7	I would imagine that most people would learn to use this dashboard very quickly.
8	I found the dashboard very complicated to use.
9	I felt very confident using the dashboard.
10	I needed to learn a lot of things before I could get going with this dashboard.

Section 3	User Experience Feedback. Please rate the following statements on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).
No.	Questions
1	The dashboard layout is clear and well-structured.
2	The charts are easy to read and interpret.
3	The colors and visuals used are appropriate and make the data understandable.
4	The dashboard provides useful insights for decision-making.
5	The navigation (tabs, sections) is intuitive.
6	Overall, I am satisfied with the dashboard experience.

Section 4	Open-Ended Feedback. Please provide your qualitative feedback in the space provided below.
No.	Questions
1	What did you like most about the dashboard?
2	What did you find confusing or difficult to use?
3	What features or improvements would you suggest?

AUTHOR'S PROFILE

Muhammad Akmal Hakim bin Ahmad Asmawi obtained Bachelor of Computer Science (Hons.) in 2024 from Universiti Teknologi MARA. Currently pursuing a Master of Information Technology at Universiti Teknologi MARA, Perak Branch, Malaysia. His research focuses on social media analytics, particularly analyzing TikTok data for business intelligence applications. He has experience in data preprocessing, natural language processing, and using AI models for sentiment analysis and topic modeling. His academic interests include text mining, digital consumer behavior, and practical applications of data science in real-world digital ecosystems.

LIST OF PUBLICATION:

Ahmad Asmawi, M. A. H., Isawasan, P., Shamugam, L., Ahmad Salleh, K., & Savita, K. S. (2025). Exploring sentiment trends in TikTok comments using GPT for influencer content strategy. e-Academia Journal, 14(1), 57-72.

Asmawi, M. A. H. A., Isawasan, P., Shamugam, L., & Salleh, K. A. (2025). A Data Science Approach to Exploring the Relationship Between TikTok Engagement and Revenue in Malaysia: A Case Study of the Beauty and Personal Care Sector. Jurnal Online Informatika, 10(2), 372-383.

Asmawi, M. A. H. A., Isawasan, P., Shamugam, L., Savita, K. S., & Budiarto, R. (2025, May). Understanding Engagement Metrics and Revenue Trends in Malaysian TikTok Shop Categories. In 2025 IEEE 15th Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 230-235). IEEE.

Isawasan, P., Asmawi, M. A. H. A., Ong, S. Q., Ooi, B. Y., & Savita, K. S. (2025, September). Comprehensive Analysis of Beauty Community Discourse on TikTok Through GPT Embeddings and BERTopic Modeling. In 2025 6th International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 444-448). IEEE.

K. S. Savita, Pradeep Isawasan, Muhammad Akmal Hakim Ahmad Asmawi, Muhammad Shaheen, Rabiya Ghafoor, "Emerging Themes and Research Directions in MOOCs and Micro-credentials", International Journal of Modern Education and Computer Science(IJMECS), Vol.17, No.6, pp. 97-110, 2025. DOI:10.5815/ijmecs.2025.06.07

Shaheen, M., Ghafoor, R., Sugathan, S. K., Isawasan, P., & Asmawi, M. A. H. A. (2026). Unveiling the Factors for MOOC Adoption: An Educational Data Mining Perspective. Information, 17(2), 175.

Tariq, R., Isawasan, P., Shamugam, L., Ahmad Asmawi, M. A. H., Nor Azman, N. A., & Zolkepli, I. A. (2025). Exploring Social Media Research Trends in Malaysia using Bibliometric Analysis and Topic Modelling. EAI Endorsed Transactions on Scalable Information Systems, 12(2).

PERPUSTAKAAN TUN ABDUL RAZAK
BAHAGIAN SUMBER RUJUKAN UNIVERSITI (BSRU)

BORANG PENYERAHAN BAHAN HARTA INTELEK UiTM

UiTM's Intellectual Property Submission Form

Nama (*Name*) **:** MUHAMMAD AKMAL HAKIM BIN AHMAD ASMAWI

Fakulti/Pusat Akademik/Bahagian : Computer and Mathematical Sciences
Faculty / Academic Centres / Department

No. Telefon (Pejabat / Hp) : +6018 288 2455
Telephone No. (Office / handphone)

E-mel (E-mail) **:** muhammadakmal2490@gmail.com

Tarikh (*Date*) **:** 11 April 2026

Pihak Fakulti / Pusat Akademik / Bahagian / Saya bersetuju bahawa dokumen dan tajuk yang disenaraikan untuk dimasukkan ke dalam Repositori Institusi UiTM.
The Faculty / Academic Centres / Department / I agree that the document and titles listed below to be placed in the UiTM Institutional Repository.

JENIS BAHAN (Sila nyatakan) : Theses
Types of Material (Please specify) :

Types of Material:
Article / Book / Theses / Bulletin / Seminar / Image / Entrepreneurship / Student Project / Research Report / Industrial Training / Annual Report / Manual / Oral History / Exam Paper / Speech / Dataset / Audio / Video / Others

MAKLUMAT BAHAN *(Information of Materials)*:

Bil. *No.*	**JUDUL BAHAN** *Title*	**HARDCOPY** *(√)*	**SOFTCOPY** *(√)*
1.	Exploring TikTok User-Generated Content For Business Intelligence Using Enhanced Data, Information, Knowledge and Wisdom (DIKW) Framework		/

**** Sila sediakan lampiran sekiranya ruangan yang disediakan tidak mencukupi*** *(Please provide attachment if necessary)*

TUJUAN PENYERAHAN BAHAN (Sila tandakan √):
Purpose (Please mark √):

1. **Bahan untuk dimuat naik ke dalam Repositori Institusi UiTM:** [/]
   (Materials for uploading into the UiTM Institutional Repository (UiTM IR)):
   Fakulti / Pusat Akademik / Bahagian / Saya perlu memastikan bahawa setiap dokumen telah dibuat semakan terlebih dahulu dan tidak mengandungi sebarang maklumat sulit sebelum diserahkan kepada pihak PTAR.
   The Faculty / Academic Centres / Department must ensure that each document has been reviewed in advance and does not contain any confidential information before being submitted to PTAR).
2. **Bahan mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan:** [ ]
   (Materials consisting of RESTRICTED information which has been determined by the organisation/body where the research was conducted):
   Sila nyatakan tarikh tamat embargo (jika ada)*:*
   (Please indicate the embargo expiry date (if any)):
   Embargo expiry date: Day: [ ] *Month:* [ ] *Year:* [ ]

   Nota: Embargo Expiry Date adalah tarikh tamat tempoh yang ditetapkan oleh penulis di mana pada atau selepas tarikh ini, bahan tersebut akan dipaparkan secara langsung di Repositori Institusi UiTM dan ianya boleh diakses.
   Note: Embargo Expiry Date *is the date that an author or a publisher imposed embargo expires. On and after this date, this document will be accessible in UiTM Institutional Repository.*
3. **Bahan boleh diakses secara teks penuh dan terbuka.** [/]
   (Materials can be accessed in full text via open access).
4. **Bahan dipinjamkan sementara untuk tujuan pendigitalan dan akan dikembalikan semula kepada pemilik.** [/]
   (Materials are on temporary loan for digitization and will be returned to the owner).

Pemberitahuan: Jika Repositori Institusi UiTM menerima bukti pelanggaran hak cipta, bahan yang berkaitan akan dikeluarkan serta-merta.
(Notification: If UiTM Institutional Repository receives proof of copyright violation, the relevant item will be removed immediately).

PERAKUAN:
Declaration:
Saya / kami akan bertanggungjawab ke atas bahan yang diserahkan untuk pendigitalan dan muat naik ke dalam Repositori Institusi UiTM.
I / we will be responsible for the materials submitted for digitization and uploaded into UiTM Institutional Repository.

Tandatangan Pemohon
Applicant Signature
Tarikh (*Date*): 11 April 2026

Tandatangan dan Cap Ketua Jabatan / Bahagian / Penyelia
Head of Division / Department / Supervisor Signature and stamp
Tarikh (*Date*): 11 April 2026

UNTUK KEGUNAAN PEJABAT (*For office use*)

DITERIMA OLEH / *Received By* :
TANDATANGAN / *Signature* :
TARIKH / *Date* :

DISAHKAN OLEH / Certified *By* :
TANDATANGAN / *Signature* :
TARIKH / *Date* :